Chapter 14
Regression

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it.

– Francis Galton (1886)

14.1 Introduction

The rather curious name regression was given to a statistical methodology by British scientist Sir Francis Galton, who analyzed the heights of sons and the average heights of their parents. From his observations, Galton concluded that sons of very tall (or short) parents were generally taller (or
shorter) than average, but not as tall (or short) as their parents. The results were published in 1886 under the title *Regression Towards Mediocrity in Hereditary Stature*. In due course of time the word *regression* became synonymous with the statistical study of the functional relationship between two or more variables. The data set illustrating Galton’s finding and used by Pearson is given in `pearson.dat`. The scatterplot and regression fits are analyzed in `galton.m` and summarized in Figure 14.1. The circles correspond to pairs of father–son heights, the black line is the line $y = x$, the red line is the regression line, and the green line is the regression line constrained to pass through the origin. Galton’s findings can be summarized by the observation that the slope of the regression (red) line was significantly smaller than the slope of the 45° line.

![Scatterplot of father–son heights](image)

*Fig. 14.1* Galton’s father–son height data (used by Pearson). The circles correspond to pairs of father–son heights, the black line is the line $y = x$, the red line is the regression line, and the green line is the regression line constrained to pass through the origin.

Usually the response variable $y$ is “regressed” on several predictors or covariates, $x_1, \ldots, x_k$, and this raises many interesting questions involving the model choice and fit, collinearity among the predictors, and others. When we have a single predictor $x$ and a linear relation between $y$ and $x$, the regression is called a simple linear regression.

### 14.2 Simple Linear Regression

Assume that we observed $n$ pairs $(x_1, y_1), \ldots, (x_n, y_n)$, and each observation $y_i$ can be modeled as a linear function of $x_i$, plus an error,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n.$$
Here $\beta_0$ and $\beta_1$ are the population intercept and slope parameters, respectively, and $\epsilon_i$ is the error. We assume that the errors are not correlated and have mean 0 and variance $\sigma^2$; thus, $\mathbb{E} y_i = \beta_0 + \beta_1 x_i$ and $\text{Var} y_i = \sigma^2$. The goal is to estimate this linear model, that is, estimate $\beta_0$, $\beta_1$, and $\sigma^2$ from the $n$ observed pairs. To put our discussion in context, we consider a study of factors affecting patterns of insulin-dependent diabetes mellitus in children.

**Example 14.1. Diabetes Mellitus in Children.** Diabetes mellitus is a condition characterized by hyperglycemia resulting from the body’s inability to use blood glucose for energy. In type 1 diabetes, the pancreas no longer makes insulin, and therefore blood glucose cannot enter the cells to be used for energy.

The objective of this study was to investigate the dependence of the level of serum C-peptide on various other factors in order to understand the patterns of residual insulin secretion. C-peptide is a protein produced by the beta cells of the pancreas whenever insulin is made. Thus, the level of C-peptide in the blood is an index of insulin production.

The part of the data from Sockett et al. (1987), discussed in the context of statistical modeling by Hastie and Tibshirani (1990), is given next. The response measurement is the logarithm of C-peptide concentration (pmol/ml) at the time of diagnosis, and the predictor is the base deficit, a measure of acidity.

<table>
<thead>
<tr>
<th>Deficit ($x$)</th>
<th>Log C-peptide ($y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-8.1</td>
<td>4.8</td>
</tr>
<tr>
<td>-16.1</td>
<td>4.1</td>
</tr>
<tr>
<td>-0.9</td>
<td>5.2</td>
</tr>
<tr>
<td>-7.8</td>
<td>5.5</td>
</tr>
<tr>
<td>-29.0</td>
<td>5</td>
</tr>
<tr>
<td>-19.2</td>
<td>3.4</td>
</tr>
<tr>
<td>-18.9</td>
<td>3.4</td>
</tr>
<tr>
<td>-10.6</td>
<td>4.9</td>
</tr>
<tr>
<td>-2.8</td>
<td>5.6</td>
</tr>
<tr>
<td>-25.0</td>
<td>3.7</td>
</tr>
<tr>
<td>-3.1</td>
<td>3.9</td>
</tr>
<tr>
<td>-9.0</td>
<td>4.9</td>
</tr>
<tr>
<td>-11.2</td>
<td>5.5</td>
</tr>
<tr>
<td>-0.2</td>
<td>4.5</td>
</tr>
<tr>
<td>-6.1</td>
<td>5.3</td>
</tr>
<tr>
<td>-1</td>
<td>4.7</td>
</tr>
<tr>
<td>-7.8</td>
<td>6.6</td>
</tr>
</tbody>
</table>

We will follow this example in MATLAB as an annotated step-by-step code/output of `cpeptide.m`. For more sophisticated analysis, MATLAB has quite advanced built-in regression tools, `regress`, `regstats`, `robustfit`, `stepwise`, and many other more-or-less specialized fitting and diagnostic tools.

After importing the data, we specify $p$, which is the number of parameters, rename the variables, and find the sample size.

```matlab
Deficit =[-8.1 -16.1 -0.9 -7.8 -29.0 -19.2 -18.9 -10.6 -2.8 ...
-25.0 -3.1 -7.8 -13.9 -4.5 -11.6 -2.1 -2.0 -9.0 -11.2 -0.2 ...
-6.1 -1 -3.6 -8.2 -0.5 -2.0 -1.6 -11.9 -0.7 -1.2 -14.3 -0.8 ...
-16.8 -5.1 -9.5 -17.0 -3.3 -3.3 -13.6 -19 -10.0 -13.5];```
It is of interest to express the log C-peptide (variable \( y \)) as a linear function of alkaline deficiency (variable \( x \)), and the population model \( y = \beta_0 + \beta_1 x + \epsilon \) is postulated. Finding estimators for \( \beta_0 \) and \( \beta_1 \) is an exercise in calculus – finding the extrema of a function of two variables. The following derivation is known as the least-squares method, which is a broad mathematical methodology for approximate solutions of overdetermined systems, first described by Gauss at the end of the eighteenth century. The best regression line minimizes the sum of squares of errors:

\[
L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.
\]

When pairs \((x_i, y_i)\) are considered fixed, \( L \) is a function of \( \beta_0 \) and \( \beta_1 \) only. Minimizing \( L \) amounts to solving the so-called normal equations

\[
\begin{align*}
\frac{\partial L}{\partial \beta_0} &= -2 \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i] = 0 \quad \text{and} \\
\frac{\partial L}{\partial \beta_1} &= -2 \sum_{i=1}^{n} [x_i y_i - \beta_0 x_i - \beta_1 x_i^2] = 0,
\end{align*}
\]

that is,

\[
\begin{align*}
n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \quad \text{and} \\
\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i.
\end{align*}
\]

(14.1)

Let \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \) and \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \) be the sample means of predictor values and the responses. If

\[
S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y},
\]

\[
S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n \bar{x}^2,
\]

\[
S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n \bar{y}^2,
\]

\[
\beta_0 = \frac{S_{xx} S_{xy} - S_{xy}^2}{S_{xx}^2 - S_{xx} S_{yy}} \\
\beta_1 = \frac{S_{yy} S_{xy} - S_{xx} S_{xy}}{S_{xx}^2 - S_{xx} S_{yy}}.
\]
14.2 Simple Linear Regression

\[ S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2, \]

and

\[ S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2, \]

then the values for \( \beta_0 \) and \( \beta_1 \) minimizing \( L \) or, equivalently, solving the normal equations (14.1) are

\[ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \]

We will simplify the notation by denoting \( \hat{\beta}_0 \) by \( b_0 \) and \( \hat{\beta}_1 \) by \( b_1 \). Thus, the fitted regression equation is

\[ \hat{y} = b_0 + b_1 x, \quad \text{with} \]

\[ b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}. \]

Note that the estimator of slope \( b_1 \) is connected with the sample correlation coefficient \( r_{XY} \),

\[ b_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \times \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} = r_{XY} \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}. \]

For values \( x = x_i \), the fitted \( \hat{y}_i \) are obtained as

\[ \hat{y}_i = b_0 + b_1 x_i, \]

with the residuals \( e_i = y_i - \hat{y}_i \). The residuals are the most important diagnostic modality in regression. They explain how well the predicted data \( \hat{y}_i \) fit the observations, and if the fit is not good, residuals indicate what caused the problem.

% Sums of Squares
SXX = sum( (x - mean(x)).^2 ) \% SXX=2.1310e+003
SYY = sum( (y - mean(y)).^2 ) \% SYY=21.807
SXY = sum( (x - mean(x)).* (y - mean(y)) ) \% SXY=105.3477
% estimators of coefficients beta1 and beta0
b1 = SXY/SXX \% 0.0494
b0 = mean(y) - b1 * mean(x) \% 5.1494
We found that $\hat{y} = 5.1494 + 0.0494x$. Figure 14.2 shows a scatterplot of log C-peptide level ($y$) against alkaline deficiency ($x$) with superimposed regression fit $\hat{y} = b_0 + b_1x$.

The sum of squared residuals, $\sum_{i=1}^n e_i^2$, is denoted by $SSE$.

One can show that $SSE = S_{yy} - b_1S_{xy}$ and that $\mathbb{E}(SSE) = (n - 2)s^2$. Thus, the mean square error $MSE = SSE/(n - 2)$ is an unbiased estimator of error variance $\sigma^2$. Recall the fundamental ANOVA identity $SST = SSR + SSE$. In regression terms, the fundamental ANOVA identity has the form

$$SST = SSR + SSE,$$

where $SST = S_{yy}$, $SSR = b_1S_{xy}$, and $SSE = \sum_{i=1}^n e_i^2$. Since

$$\mathbb{E}SSR = \sigma^2 + \beta_1^2S_{xx},$$

$SSR$ has an associated 1 degree of freedom and the regression mean sum of squares $MSR$ is $SSR/1 = SSR$.

The statistic $MSR$ becomes an unbiased estimator of variance $\sigma^2$ when $\beta_1 = 0$. Thus, if $H_0 : \beta_1 = 0$ is true, one should have $F = MSR/MSE$ close to 1, since under $H_0$ both $MSR$ and $MSE$ estimate the same quantity, $\sigma^2$. Under $H_0$, the statistic $F = MSR/MSE$ has an $F$-distribution with 1 and $n - 2$ degrees of freedom.

Large values of $F$ indicate that there is a contribution of $\beta_1$ in $MSR$, and discrepancy from $H_0$ can be assessed using an $F$-test. The sums of squares,
degrees of freedom, mean squares, F-statistic and p-value associated with observed F are customarily summarized in an ANOVA table:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSR</td>
<td>MSR</td>
<td>( F = \frac{SSR}{MSE} ) ( P(F_{1,n-2} &gt; F) )</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>( n - 2 )</td>
<td>SSE</td>
<td>MSE</td>
<td>( MSE = \frac{SSE}{n-2} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n - 1 )</td>
<td>SST</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The p-value is associated with testing of \( H_0 \), which essentially states that covariate \( x \) does not influence the response \( y \), and the same fit can be obtained by just taking \( y \) as the model for \( y_i \).

The preceding calculations are arranged in the ANOVA table:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>5.2079</td>
<td>5.2079</td>
<td>12.8637</td>
<td>0.0009</td>
</tr>
<tr>
<td>Error</td>
<td>( n - 2 )</td>
<td>16.5990</td>
<td>0.4049</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n - 1 )</td>
<td>21.8070</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 14.3a shows a plot of residuals \( y_i - \hat{y}_i \) against \( x_i \). A normalized histogram of residuals with a superimposed normal distribution \( \mathcal{N}(0,0.6363^2) \), is given in Figure 14.3b.

The quantity \( R^2 \), called the coefficient of determination, is defined as

\[
R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.
\]

The \( R^2 \) in this context coincides with the square of the correlation coefficient between \( (x_1,\ldots,x_n) \) and \( (y_1,\ldots,y_n) \). However, the representation of \( R^2 \) via the ratio SSR/SST is more illuminating. In words, \( R^2 \) explains what proportion of the total variability (SST) encountered in observations is explained or accounted for by the regression (SSR). Thus, a high \( R^2 \) is desirable in any regression. Note that \( F = \frac{MSR}{MSE} = \frac{(n-2)R^2}{1-R^2} \) and the F test is
equivalent to testing that the population correlation coefficient ρ (the correlation between X’s and Y’s) is significant.

The adjusted $R^2$ (Ezekiel, 1930), defined as

$$R^2_{adj} = 1 - \frac{n - 1}{n - p} \frac{SSE}{SST} = 1 - \frac{MSE}{MST},$$

where $MST = SST/(n - 1) = s^2_y$, is important in cases with multiple predictors ($p > 2$), since it penalizes inclusion of predictors in the model.

Often, instead of regressing $y_i$ on $x_i$, one regresses $y_i$ on $x_i - \bar{x}$ as

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}).$$

This is beneficial for several reasons. In practice, we calculate only the estimator $b_1$. Since the fitted line contains the point $(\bar{x}, \bar{y})$, the intercept $\beta_0^*$ is estimated by $\bar{y}$, and our regression fit is

$$\hat{y}_i = \bar{y} + b_1(x_i - \bar{x}).$$

In the Bayesian context, estimating $\beta_0^*$ and $\beta_1$ is more stable and efficient than estimating $\beta_0$ and $\beta_1$ directly, since $\bar{y}$ and $b_1$ are uncorrelated.

Estimators $b_0$ and $b_1$ are unbiased estimators of population’s $\beta_0$ and $\beta_1$. We will show that they are unbiased and that their variance is intimately connected with the variance of responses, $\sigma^2$.
\( \mathbb{E}b_1 = \beta_1 \) and \( \text{Var} b_1 = \sigma^2 / S_{xx} \).

\( \mathbb{E}b_0 = \beta_0 \) and \( \text{Var} b_0 = \sigma^2 \left( \frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}} \right) \).

Here is the rationale:

\[
\mathbb{E}b_1 = \mathbb{E} \left( \frac{S_{xy}}{S_{xx}} \right) = \frac{1}{S_{xx}} \sum_{i=1}^{n} y_i (x_i - \bar{x}) \\
= \frac{1}{S_{xx}} \mathbb{E} \left( \sum_{i=1}^{n} (\beta_0^* + \beta_1 (x_i - \bar{x}) + \epsilon_i)(x_i - \bar{x}) \right) \\
= \frac{1}{S_{xx}} \left[ \sum_{i=1}^{n} \beta_0^* (x_i - \bar{x}) + \sum_{i=1}^{n} \beta_1 (x_i - \bar{x})^2 + \mathbb{E} \sum_{i=1}^{n} \epsilon_i (x_i - \bar{x}) \right] \\
= \frac{1}{S_{xx}} [0 + \beta_1 S_{xx} + 0] = \beta_1.
\]

\[
\text{Var} b_1 = \text{Var} \left( \frac{S_{xy}}{S_{xx}} \right) = \frac{1}{S_{xx}^2} \sum_{i=1}^{n} \text{Var} (y_i (x_i - \bar{x})) \\
= \frac{1}{S_{xx}^2} \sum_{i=1}^{n} \sigma^2 (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}}.
\]

Since \( b_0 = \bar{y} - b_1 \bar{x} \),

\[
\mathbb{E}b_0 = \mathbb{E}(\bar{y} - b_1 \bar{x}) = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0
\]

and

\[
\text{Var} b_0 = \text{Var}(\bar{y}) + \text{Var}(b_1 \bar{x}) - 2 \text{Cov}(\bar{y}, b_1 \bar{x}) \\
= \frac{\sigma^2}{n} + (\bar{x})^2 \frac{\sigma^2}{S_{xx}} - 2 \cdot 0 = \sigma^2 \left[ \frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}} \right].
\]

An alternative expression for \( \text{Var} b_0 \) is \( \sigma^2 \frac{\bar{x}^2}{S_{xx}} \), for \( \bar{x}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \). Sample counterparts of \( \text{Var} b_0 \) and \( \text{Var} b_1 \) will be needed for the inference in subsequent sections; they are obtained by plugging in the \( \text{MSE} \) in place of \( \sigma^2 \).

The covariance between \( b_0 \) and \( b_1 \) is
\[ \text{Cov}(b_0, b_1) = \text{Cov}(\overline{y} - b_1 \cdot \overline{x}, b_1) = \text{Cov}(\overline{y}, b_1) - \overline{x} \cdot \text{Var}(b_1) = -\overline{x} \cdot \frac{\sigma^2}{S_{xx}}, \]

since \( \text{Cov}(\overline{y}, b_1) = 0 \). The correlation between \( b_0 \) and \( b_1 \) is then readily found as

\[ \text{Corr}(b_0, b_1) = -\frac{\overline{x}}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}}. \]

In MATLAB the estimator of \( \sigma \) and sample standard deviations of estimators \( b_0 \) and \( b_1 \) from Example 14.1 are as follows:

```matlab
% s, sb0, and sb1
s = sqrt(MSE) %s = 0.6363
%Standard errors of parameter estimators
sb1 = s/sqrt(SXX) %sb1 = 0.0138
sb0 = s * sqrt(1/n + (mean(x))^2/SXX) %sb0 = 0.1484
```

### 14.3 Inference in Simple Linear Regression

To find the estimators of regression parameters and calculate their expectations and variances, we do not need the distributional properties of errors, except that they are independent, have a mean 0, and a variance that does not vary with \( x \). However, to test the hypotheses about the population intercept and slope, and to find confidence intervals, we need to assume that the errors \( \epsilon_i \) are i.i.d. normal. In practice, the residual analysis is conducted to verify whether the normality assumption is justified.

#### 14.3.1 Inference about the Slope Parameter

For a given constant \( \beta_{10} \), the test for

\[ H_0: \beta_1 = \beta_{10} \]

relies on the statistic

\[ t = \frac{b_1 - \beta_{10}}{s^2/S_{xx}}. \]

where \( s^2 = \text{MSE} \). This statistic under \( H_0 \) has a \( t \)-distribution with \( n - 2 \) degrees of freedom, and testing is done as follows:
14.3 Inference in Simple Linear Regression

<table>
<thead>
<tr>
<th>Alternative</th>
<th>$a$-level rejection region</th>
<th>$p$-value (MATLAB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1 : \beta_1 &gt; \beta_{10}$</td>
<td>$[t_{n-2,1-a,\infty})$</td>
<td>1-\text{tcdf}(t,n-2)</td>
</tr>
<tr>
<td>$H_1 : \beta_1 \neq \beta_{10}$</td>
<td>$(-\infty,t_{n-2,a/2}] \cup [t_{n-2,1-a/2,\infty})$</td>
<td>2*\text{tcdf}(-\text{abs}(t),n-2)</td>
</tr>
<tr>
<td>$H_1 : \beta_1 &lt; \beta_{10}$</td>
<td>$(-\infty,t_{n-2,a})$</td>
<td>\text{tcdf}(t,n-2)</td>
</tr>
</tbody>
</table>

The distribution of the test statistic is derived from a linear representation of $b_1$ as

$$b_1 = \sum_{i=1}^{n} a_i y_i, \quad a_i = \frac{x_i - \bar{x}}{S_{xx}}.$$ 

Under $H_0$, $b_1 \sim \mathcal{N}(\beta_{10}, \sigma^2 / S_{xx})$. Thus,

$$t = \frac{b_1 - \beta_{10}}{\sqrt{s^2 / S_{xx}}} = \frac{b_1 - \beta_{10}}{\sigma \sqrt{s^2 / S_{xx}}} \times \frac{s}{\sqrt{\frac{SSE}{n-2} \sigma^2}},$$

which, by definition, has a $t_{n-2}$-distribution, as $Z / \sqrt{\frac{1}{n-2} \frac{s^2}{\sigma^2}}$. We also used the fact that $s^2 = MSE = SSE / (n - 2)$.

The $(1 - a)100\%$ confidence interval for $\beta_1$ is

$$[b_1 - t_{n-2,1-a/2} \frac{s}{\sqrt{S_{xx}}}, \quad b_1 + t_{n-2,1-a/2} \frac{s}{\sqrt{S_{xx}}}]$$.

14.3.2 Inference about the Intercept Parameter

For a given constant $\beta_{00}$, the test for

$$H_0 : \beta_0 = \beta_{00}$$

relied on the statistic

$$t = \frac{b_0 - \beta_{00}}{s \sqrt{\frac{1}{2} + \frac{(\bar{y})^2}{S_{xx}}}}.$$
Under $H_0$ this statistic has a $t$-distribution with $n - 2$ degrees of freedom, and testing is done as follows:

<table>
<thead>
<tr>
<th>Alternative</th>
<th>$a$-level rejection region</th>
<th>$p$-value (MATLAB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1 : \beta_0 &gt; \beta_{00}$</td>
<td>$[t_{n-2,1-a,\infty}]$</td>
<td>$1-\text{tcdf}(t,n-2)$</td>
</tr>
<tr>
<td>$H_1 : \beta_0 \neq \beta_{00}$</td>
<td>$(-\infty,t_{n-2,a/2}] \cup [t_{n-2,1-a/2,\infty})$</td>
<td>$2*\text{tcdf}(-\text{abs}(t),n-2)$</td>
</tr>
<tr>
<td>$H_1 : \beta_0 &lt; \beta_{00}$</td>
<td>$(-\infty,t_{n-2,a}]$</td>
<td>$\text{tcdf}(t,n-2)$</td>
</tr>
</tbody>
</table>

This is based on the representation of $b_0$ as $b_0 = \overline{y} - b_1\overline{x}$ and under $H_0$ $b_0 \sim \mathcal{N}(\beta_{00},\sigma^2\left(\frac{1}{n} + (\overline{x})^2/S_{xx}\right))$. Thus,

$$t = \frac{b_0 - \beta_{00}}{s\sqrt{1/n + (\overline{x})^2/S_{xx}}} = \frac{b_1 - \beta_{00}}{\sigma\sqrt{1/n + (\overline{x})^2/S_{xx}}} \times \frac{\sigma}{s} = \frac{b_0 - \beta_{00}}{\sigma\sqrt{1/n + (\overline{x})^2/S_{xx}}} \times \frac{s}{\sqrt{S_{SE}(n-2)\sigma^2}},$$

which by definition has a $t_{n-2}$-distribution, as $Z/\sqrt{\chi^2_{n-2}}$.

The $(1 - \alpha)100\%$ confidence interval for $\beta_0$ is

$$\left[b_0 - t_{n-2,1-a/2} s\sqrt{\frac{1}{n} + (\overline{x})^2/S_{xx}}, b_0 + t_{n-2,1-a/2} s\sqrt{\frac{1}{n} + (\overline{x})^2/S_{xx}}\right].$$

% Are the coefficients equal to 0?
\begin{verbatim}
t1 = b1/sb1 \%3.5866
pbl1 = 2 * (1 - tcdf(abs(t1),n-p) ) \%8.8412e-004
t0 = b0/sb0 \%34.6927
pb0 = 2 * (1 - tcdf(abs(t0),n-p) ) \%0
\end{verbatim}

% Test H_0: beta1 = 0.04 vs. H_1: beta1 > 0.04
tst1 = (b1 - 0.04)/sb1 \%0.6846
ptst1 = 1 - tcdf( tst1, n-p ) \%0.2487

% Test H_0: beta0 = 5.8 vs. H_1: beta0 < 5.8
tst2 = (b0 - 5.8)/sb0 \%-4.3836
ptst2 = tcdf(tst2, n-p ) \%3.9668e-005

% Find 95% CI for beta1
\begin{verbatim}
[b1 - tinv(0.975, n-p)*sb1, b1 + tinv(0.975, n-p)*sb1]
\end{verbatim}
% 0.0216 0.0773

% Find 99% CI for beta0
\begin{verbatim}
[b0 - tinv(0.995, n-p)*sb0, b0 + tinv(0.995, n-p)*sb0]
\end{verbatim}
% 4.7484 5.5503
14.3 Inference in Simple Linear Regression

14.3.3 Inference about the Variance

Testing \( H_0 : \sigma^2 = \sigma_0^2 \) relies on the statistic \( \chi^2 = \frac{(n-2)\text{MSE}}{\sigma_0^2} = \frac{\text{SSE}}{\sigma_0^2} \). This statistic under \( H_0 \) has a \( \chi^2 \)-distribution with \( n-2 \) degrees of freedom and testing is done as follows:

<table>
<thead>
<tr>
<th>Alternative</th>
<th>( \alpha )-level rejection region</th>
<th>( p )-value (MATLAB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_1 : \sigma^2 &lt; \sigma_0^2 )</td>
<td>( [0, \chi^2_{n-2, a/2}] )</td>
<td>\text{chi2cdf(chi2,n-2)}</td>
</tr>
<tr>
<td>( H_1 : \sigma^2 \neq \sigma_0^2 )</td>
<td>( [0, \chi^2_{n-2, a/2}] \cup [\chi^2_{n-2, 1-a/2}, \infty) )</td>
<td>( 2 \times \text{chi2cdf(ch,n-2)} )</td>
</tr>
<tr>
<td>( H_1 : \sigma^2 &gt; \sigma_0^2 )</td>
<td>( [\chi^2_{n-2, 1-a/2}, \infty) )</td>
<td>( 1 - \text{chi2cdf(chi2,n-2)} )</td>
</tr>
</tbody>
</table>

where \( \chi2 \) is the test statistic and \( ch = \min(chi2, 1/\chi2) \).

The \((1 - \alpha)100\%\) confidence interval for \( \sigma^2 \) is

\[
\left[ \frac{\text{SSE}}{\chi^2_{n-2-1-a/2}}, \frac{\text{SSE}}{\chi^2_{n-2-a/2}} \right].
\]

The following MATLAB script tests \( H_0 : \sigma^2 = 0.5 \) versus \( H_1 : \sigma^2 < 0.5 \) and finds a 95% confidence interval for \( \sigma^2 \). As is evident, \( H_0 \) is not rejected (\( p \)-value 0.1981), and the interval is \([0.2741, 0.6583]\).

```matlab
% Test H0: sigma2 = 0.5 vs. H1: sigma2 < 0.5
ch2 = SSE/0.5  \% 33.1981
ptst3 = chi2cdf(ch2, n-p) \% 0.1981
% Find 95% CI for sigma2
[SSE/ch2inv(0.975, n-p), SSE/ch2inv(0.025, n-p)]
% 0.2741 0.6583
```

14.3.4 Inference about the Mean Response

Suppose that the regression \( \hat{y} = b_0 + b_1 x \) has been found and that we are interested in making an inference about the response \( y_m = E(y|x = x^*) = \beta_0 + \beta_1 x^* \). The statistic for \( y_m \) is \( \hat{y}_m = b_0 + b_1 x^* \), and it is a random variable since both \( b_0 \) and \( b_1 \) are random variables.

The \( \hat{y}_m \) is an unbiased estimator of \( y_m \), \( E(\hat{y}_m) = E(b_0 + b_1 x^*) = \beta_0 + \beta_1 x^* = y_m \), as expected. The variance of \( \hat{y}_m \) is obtained from representation \( \hat{y}_m = b_0 + b_1 x^* = \overline{y} + b_1 (x^* - \overline{x}) \) and the fact that the correlation between \( \overline{y} \) and \( b_1 \) is zero:
\[ \text{Var} \hat{y}_m = \sigma^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right). \]

Thus,

\[ \hat{y}_m \sim \mathcal{N} \left( \beta_0 + \beta_1 x^*, \sigma^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right) \right), \]

from which we develop the inference.

The test

\[ H_0 : y_m = y_0 \]

relies on the statistic

\[ t = \frac{\hat{y}_m - y_0}{s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}. \]

This statistic under \( H_0 \) has a \( t \)-distribution with \( n - 2 \) degrees of freedom and testing is done as in the cases of \( \beta_0 \) and \( \beta_1 \).

The \((1 - \alpha)100\%\) confidence interval for \( y_m = \beta_0 + \beta_1 x^* \) is

\[ \left[ \hat{y}_m - t_{n, 1 - \alpha/2} s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \ , \ \hat{y}_m + t_{n, 1 - \alpha/2} s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \right]. \]

### 14.3.5 Inference about a New Response

Suppose that the regression \( \hat{y} = b_0 + b_1 x \) has been established and that we are interested in predicting the response \( \hat{y}_{\text{pred}} \) for a new observation, corresponding to a covariate \( x = x^* \). Given the value \( x = x^* \), the difference between the inference about the mean response \( y_m \) discussed in the previous section and the inference about an individual outcome \( y_{\text{pred}} \) is substantial.

As in the previous subsection, \( \hat{y}_{\text{pred}} = b_0 + b_1 x^* \), and the mean of \( \hat{y}_{\text{pred}} \) is \( \mathbb{E}(\hat{y}_{\text{pred}}) = \beta_0 + \beta_1 x^* = y_{\text{pred}} \), which is in fact equal to \( y_m \).

Where \( y_{\text{pred}} \) and \( y_m \) differ is in their variability. The variability of \( \hat{y}_{\text{pred}} \) has two sources, first, the variance of the distribution of \( y_0 \) for \( x = x^* \), which is \( \sigma^2 \), and, second, the variance of sampling distribution for \( b_0 + b_1 x^* \), which is \( \sigma^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right) \). Thus, \( \text{Var}(\hat{y}_{\text{pred}}) = \text{MSE} + \text{Var}(\hat{y}_m) \).

The distribution for \( \hat{y}_{\text{pred}} \) is normal,
\[ \hat{y}_{\text{pred}} \sim \mathcal{N}(\beta_0 + \beta_1 x^*, \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)) , \]

and the subsequent inference is based on this distribution.

The test

\[ H_0 : y_{\text{pred}} = y_0 \]

relies on the statistic

\[ t = \frac{\hat{y}_{\text{pred}} - y_0}{s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}. \]

This statistic under \( H_0 \) has a \( t \)-distribution with \( n - 2 \) degrees of freedom, which implies the inference.

The \((1 - \alpha)100\%\) confidence interval for \( y_{\text{pred}} \) is

\[ \left[ \hat{y}_{\text{pred}} - t_{n-2,1 - \alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} , \hat{y}_{\text{pred}} + t_{n-2,1 - \alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \right]. \]

Remark. Suppose that for \( x = x^* \), instead of a single new response, we anticipate \( m \) new responses and wish to find the prediction interval for
their average. The prediction interval in this case is obtained by replacing \( \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \) with \( \sqrt{1 + \frac{1}{m} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \).

Next, we will find Bayesian estimators of the regression parameters in the same example, *Diabetes Mellitus in Children*, by using WinBUGS. On page 688 we mentioned that taking \( x_i - \bar{x} \) as a predictor instead of \( x_i \) is beneficial in the Bayesian context. From such a parametrization of regression,

\[
y_i = \beta_0^* + \beta_1 (x_i - \bar{x}) + \epsilon_i,
\]

the traditional intercept \( \beta_0 \) is then obtained as \( \beta_0^* - \beta_1 \bar{x} \).

\[
\text{model}\{
  \text{for (i in 1:ntotal)}{
    y[i] ~ dnorm( mui[i], tau )
    mui[i] <- bb.0 + b.1 *(x[i] - mean(x[]))
    yres[i] <- y[i] - mui[i]
  }
  bb.0 ~ dnorm(0, 0.0001)
  b.0 <- bb.0 - b.1 * mean(x[])
  b.1 ~ dnorm(0, 0.0001)
  tau ~ dgamma(0.001, 0.001)
  s <- 1/sqrt(tau)
\}

\[
\text{DATA}\n\text{list(ntotal=43,}
\text{  y = c(4.8, 4.1, 5.2, 5.5, 5.0, 3.4, 3.4, 4.9, 5.6, 3.7,}
\text{  3.9, 4.5, 4.8, 4.9, 3.0, 4.6, 4.8, 5.5, 4.5, 5.3,}
\text{  4.7, 6.6, 5.1, 3.9, 5.7, 5.1, 5.2, 3.7, 4.9, 4.8,}
\text{  4.4, 5.2, 5.1, 4.6, 3.9, 5.1, 5.1, 6.0, 4.9, 4.1,}
\text{  4.6, 4.9, 5.1),}
\text{  x = c(-8.1, -16.1, -0.9, -7.8, -29.0, -19.2, -18.9, -10.6,}
\text{  -2.8, -25.0, -3.1, -7.8, -13.9, -4.5, -11.6, -2.1,}
\text{  -2.0, -9.0, -11.2, -0.2, -6.1, -1.0, -3.6, -8.2,}
\text{  -0.5, -2.0, -1.6, -11.9, -0.7, -1.2, -14.3, -0.8,}
\text{  -16.8, -5.1, -9.5, -17.0, -3.3, -0.7, -3.3, -13.6,}
\text{  -1.9, -10.0, -13.5))}
\]

\[
\text{INITS}\n\text{list(bb.0 = 0, b.1 = 0, tau=1)}
\]

The output is given in the table below. It contains Bayesian estimators \( b_0 \) for \( \beta_0 \) and \( b_1 \) for \( \beta_1 \). In the least-squares regression we found that \( b_1 = S_{xy}/S_{xx} = 0.0494, b_0 = \bar{y} - b_1 \cdot \bar{x} = 5.1494, \) and \( s = \sqrt{MSE} = 0.6363 \). Since priors were noninformative, we expect that the Bayes estimators will be close to the classical. Indeed that is the case: \( b_0 = 5.149, b_1 = 0.0494, \) and \( s = 0.6481 \).
The classical standard errors of estimators for $\beta_0$ and $\beta_1$ are $sb_0 = 0.1484$ and $sb_1 = 0.0138$, while the corresponding Bayesian estimators are 0.1525 and 0.01418. The classical 95% confidence interval for $\beta_1$ was found to be $[0.0216, 0.0773]$. The Bayesian 95% credible set for $\beta_1$ is $[0.02139, 0.07733]$, as is evident from $val_{2.5pc}$ and $val_{97.5pc}$ in the output below:

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>val2.5pc</th>
<th>median</th>
<th>val97.5pc</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>b.0</td>
<td>5.149</td>
<td>0.1325</td>
<td>3.117E-4</td>
<td>4.848</td>
<td>5.149</td>
<td>5.449</td>
<td>2001</td>
<td>200000</td>
</tr>
<tr>
<td>b.1</td>
<td>0.0494</td>
<td>0.0141</td>
<td>3.072E-5</td>
<td>0.02139</td>
<td>0.04944</td>
<td>0.07733</td>
<td>2001</td>
<td>200000</td>
</tr>
<tr>
<td>s</td>
<td>0.6481</td>
<td>0.0734</td>
<td>1.771E-4</td>
<td>0.5236</td>
<td>0.6415</td>
<td>0.811</td>
<td>2001</td>
<td>200000</td>
</tr>
<tr>
<td>yres[1]</td>
<td>0.05111</td>
<td>0.09944</td>
<td>2.175E-4</td>
<td>-0.1444</td>
<td>0.05125</td>
<td>0.2472</td>
<td>2001</td>
<td>200000</td>
</tr>
<tr>
<td>yres[2]</td>
<td>-0.2537</td>
<td>0.1502</td>
<td>3.459E-4</td>
<td>-0.5499</td>
<td>-0.2533</td>
<td>0.0418</td>
<td>2001</td>
<td>200000</td>
</tr>
<tr>
<td>yres[3]</td>
<td>0.09544</td>
<td>0.1431</td>
<td>2.925E-4</td>
<td>-0.1861</td>
<td>0.09505</td>
<td>0.376</td>
<td>2001</td>
<td>200000</td>
</tr>
<tr>
<td>yres[4]</td>
<td>0.7363</td>
<td>0.09957</td>
<td>2.167E-4</td>
<td>0.5406</td>
<td>0.7364</td>
<td>0.9525</td>
<td>2001</td>
<td>200000</td>
</tr>
<tr>
<td></td>
<td>yres[41]</td>
<td>-0.4552</td>
<td>0.1333</td>
<td>2.727E-4</td>
<td>-0.7173</td>
<td>-0.4555</td>
<td>-0.1919</td>
<td>2001</td>
</tr>
<tr>
<td></td>
<td>yres[42]</td>
<td>0.245</td>
<td>0.1028</td>
<td>2.314E-4</td>
<td>0.04251</td>
<td>0.2451</td>
<td>0.4475</td>
<td>2001</td>
</tr>
<tr>
<td></td>
<td>yres[43]</td>
<td>0.6179</td>
<td>0.125</td>
<td>2.879E-4</td>
<td>0.3718</td>
<td>0.6179</td>
<td>0.8632</td>
<td>2001</td>
</tr>
</tbody>
</table>

Thus, the Bayesian approach to regression estimation is quite close to the classical when the priors on $\beta_0$ and $\beta_1$ and the precision $\tau = 1/\sigma^2$ are noninformative.

**Example 14.2. Hubble Regression.** Hubble’s constant (H) is one of the most important numbers in cosmology because it is instrumental in estimating the size and age of the universe. This long-sought number indicates the rate at which the universe is expanding, from the primordial “Big Bang.” The Hubble constant can be used to determine the intrinsic brightness and masses of stars in nearby galaxies, examine those same properties in more distant galaxies and galaxy clusters, deduce the amount of dark matter present in the universe, obtain the scale size of faraway galaxy clusters, and serve as a test for theoretical cosmological models.

In 1929, Edwin Hubble, a distinguished American astronomer, investigated the relationship between the distance of a galaxy from the Earth and the velocity with which it appears to be receding. Galaxies appear to be moving away from us no matter which direction we look. This is thought to be the result of the “Big Bang.” Hubble hoped to provide some knowledge about how the universe was formed and what might happen in the future. The data collected included distances (megaparsecs1) to $n = 24$ galaxies and their recessional velocities (km/sec).

Hubble’s law is as follows: Recessional velocity = $H \times$ distance, where $H$ is Hubble’s constant (units of $H$ are [km/sec/Mpc]). By working backwards in time, the galaxies appear to meet in the same place. Thus $1/H$ can be used to estimate the time since the Big Bang, a measure of the age of the universe.

---

1 1 parsec = 3.26 light years
A regression analysis seems appropriate; however, there is no intercept term in Hubble’s law. Can you verify that the constant term of the regression analysis is not significantly different than 0 at any reasonable level of $\alpha$. Find the 95% confidence interval for the slope $\beta_1$, also known as Hubble’s constant $H$, from the given data.

The age of the universe as predicted by Hubble (in years) is about 2.3 billion years.

Fig. 14.4 Hubble’s data and regression fits. The blue line is an unconstrained regression (with intercept fitted), and the red line is a no-intercept fit. The slope for the no-intercept fit is $b_1 = 423.9373 (=H)$.

$$H = 423.9373$$

secinyear =60+60+24+365 \%31536000
kminmps = 3.08568025 * 10^19;
age = 1/H * kminmps/secinyear \%2.3080e+009

Modern measurements put $H$ at approx. 70, thus predicting the age of the universe to about 14 billion years. Figure 14.4 showing Hubble’s data and regression fits is generated by hubble.m.

2 Reasonable here means level $\alpha$ not larger than 0.10.
Often in regression problems we need to make an inference about the predictor \( x \) when a response \( y \) is observed or assumed. This typically arises in the context of instrument calibration, which gives the name to the statistical methodology used to solve this kind of problems.

A naïve solution, sometimes referred as the reverse method, is to reverse the roles of \( x \) and \( y \), fit a regression \( \hat{x} = c_0 + c_1 y \), and apply the results from Section 14.3.5. The problem with this approach is that it assumes that measurements \( y_i \) are observed without error, while \( x_i \)'s are observed with error, which may not be the case since \( x \)'s are typically controlled. More seriously, when \( y_i \)'s are random, the values \( y^*, \bar{y}, \) and \( S_{yy} \) in

\[
s_{x0} = s_x \sqrt{1 + \frac{1}{n} + \frac{(y^* - \bar{y})^2}{S_{yy}}}, \quad s_x = \sqrt{\frac{S_{xx}}{n - 2}},
\]

are not fixed constants but random, unlike \( x^*, \bar{x}, \) and \( S_{xx} \) in a reverse counterpart \( s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \) from Section 14.3.5.

The standard method uses the original regression \( \hat{y} = b_0 + b_1 x \) and, for the response \( y^* \), predicts \( x^* \) as

\[
\hat{x}^* = \frac{1}{b_1} (y^* - b_0).
\]

This method is called inverse method because the linear equation is inverted, In this case the expectation and variance of \( \hat{x}^* \) are approximated as

\[
E\hat{x}^* \approx x^* + \frac{\sigma^2}{\beta_1^2 S_{xx}} (x^* - \bar{x}), \quad \text{Var} \hat{x}^* \approx \frac{\sigma^2}{\beta_1^2} \left( 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right).
\]

The \((1 - \alpha)100\%\) prediction interval for \( x^* \) is

\[
\left[ \hat{x}^* - t_{n-2,1-\alpha/2} \frac{s}{\beta_1} \sqrt{1 + \frac{1}{n} + \frac{(\hat{x}^* - \bar{x})^2}{S_{xx}}}, \hat{x}^* + t_{n-2,1-\alpha/2} \frac{s}{\beta_1} \sqrt{1 + \frac{1}{n} + \frac{(\hat{x}^* - \bar{x})^2}{S_{xx}}} \right].
\]

If for a single \( x^*, y_1^*, y_2^*, \ldots, y_m^* \) are observed, then \( y^* \) from above is replaced by the mean \( \bar{y}^* \), \( \hat{x}^* = \frac{1}{m_1} (\bar{y}^* - b_0) \), and the prediction interval becomes
Example 14.3. Concentration of Caprolactone. Thonnard (2006) analyzes data on 10 solutions of caprolactone in the solvent tetrahydrofuran. As a control, a solution without the caprolactone is also provided.

Each of these 11 solutions was injected in a gas chromatograph three times. The measures from each injection are recorded in a form of surface, which translates to the estimated concentration. The known solution concentrations \( x \) for the 10 solutions are paired with the surface readings \( y \), three for each concentration. The no caprolactone solution results in three pairs \((0, 0)\). Thus, the data presented in Table 14.1 consist of 33 \((x, y)\) pairs:

<table>
<thead>
<tr>
<th>Observations</th>
<th>Concentration x (in g/l)</th>
<th>Surface Measures y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–3</td>
<td>9.71</td>
<td>24.276, 24.083, 24.276</td>
</tr>
<tr>
<td>4–6</td>
<td>8.52</td>
<td>20.206, 20.199, 20.223</td>
</tr>
<tr>
<td>7–9</td>
<td>7.96</td>
<td>19.773, 19.759, 19.765</td>
</tr>
<tr>
<td>10–12</td>
<td>6.82</td>
<td>16.743, 16.587, 16.744</td>
</tr>
<tr>
<td>13–15</td>
<td>5.85</td>
<td>15.081, 15.121, 15.274</td>
</tr>
<tr>
<td>16–18</td>
<td>4.95</td>
<td>12.636, 12.641, 12.682</td>
</tr>
<tr>
<td>22–24</td>
<td>2.98</td>
<td>7.624, 7.592, 7.585</td>
</tr>
<tr>
<td>25–27</td>
<td>2.07</td>
<td>4.638, 4.666, 4.649</td>
</tr>
<tr>
<td>28–30</td>
<td>1.02</td>
<td>2.860, 2.859, 2.896</td>
</tr>
<tr>
<td>31–33</td>
<td>0.00</td>
<td>0.00, 0.00, 0.00</td>
</tr>
</tbody>
</table>

Table 14.1 Concentration and surface of caprolactone in the solution.

Suppose that we have a new solution, for which we do not know the concentration of caprolactone, \( x^* \). After injecting this unknown solution three times in the gas chromatograph, three observations of \( y \) are obtained, \( y_1^* = 1.582, y_2^* = 1.793, \) and \( y_3^* = 1.787 \). We will estimate \( x^* \) and provide the 95% prediction interval.

MATLAB script `caprolactone.m` finds the prediction \( \hat{x}^* = 0.6329 \) and the 95% prediction interval for \( x^* \) as \([0.4047,0.8611]\).
14.5 Testing the Equality of Two Slopes*

Let \((x_{1i}, y_{1i}), \ i = 1, \ldots, n_1\) and \((x_{2i}, y_{2i}), \ i = 1, \ldots, n_2\), be pairs of measurements obtained from two groups, and for each group the regression is estimated as

\[
\begin{align*}
y_{1i} &= b_{0(1)} + b_{1(1)}x_{1i} + e_{1i}, \ i = 1, \ldots, n_1, \text{ and} \\
y_{2i} &= b_{0(2)} + b_{1(2)}x_{2i} + e_{2i}, \ i = 1, \ldots, n_2,
\end{align*}
\]

where in groups \(i = 1, 2\) the statistics \(b_{0(i)}\) and \(b_{1(i)}\) are estimators of the respective population parameters, intercepts \(\beta_{0(i)}\) and slopes \(\beta_{1(i)}\). We are interested in testing the equality of the population slopes,

\[H_0 : \beta_{1(1)} = \beta_{1(2)},\]

against the one- or two-sided alternatives.

The test statistic is

\[
t = \frac{b_{1(1)} - b_{1(2)}}{s.e.(b_{1(1)} - b_{1(2)})},
\]

(14.2)
where the standard error of the difference \( b_{1(1)} - b_{1(2)} \) is

\[
s.e.(b_{1(1)} - b_{1(2)}) = \sqrt{s^2 \left[ \frac{1}{S_{xx(1)}} + \frac{1}{S_{xx(2)}} \right]},
\]

and \( s^2 \) is the pooled estimator of variance,

\[
s^2 = \frac{SSE_1 + SSE_2}{n_1 + n_2 - 4}.
\]

Statistic \( t \) in (14.2) has a \( t \)-distribution with \( n_1 + n_2 - 4 \) degrees of freedom and, in addition to testing, could be used for a \( (1 - \alpha) \) 100\% confidence interval for \( \hat{\beta}_{1(1)} - \hat{\beta}_{1(2)} \),

\[
[(b_{1(1)} - b_{1(2)}) \pm t_{n_1+n_2-4,1-\alpha/2} \times s.e.(b_{1(1)} - b_{1(2)})].
\]

**Example 14.4. Cadmium Poisoning.** Chronic cadmium poisoning is an insidious disease associated with the development of emphysema and the excretion in the urine of a characteristic protein of low molecular weight. The first signs of chronic cadmium poisoning become apparent following a latent interval after exposure has ended. Respiratory functions deteriorate faster with age. The data set featured in Armitage and Berry (1994) gives ages (in years) and vital capacity (in liters) for 84 men working in the cadmium industry, `cadmium.dat`|mat|xlsx. The observations with flag `exposure` equal to 0 denote persons unexposed to cadmium oxide fumes, while flags 1 and 2 correspond to exposed persons. The purpose of the study was to assess the degree of influence of exposure to respiratory functions. Since respiratory functions are influenced by age, regardless of exposure, age as a covariate needs to be taken into account. Thus, the suggested methodology is to test the equality of the slopes in group regressions of vital capacity to age:

\[
H_0 : \beta_{1(exposed)} = \beta_{1(unexposed)} \quad \text{versus} \quad H_1 : \beta_{1(exposed)} < \beta_{1(unexposed)}.
\]

The research hypothesis is that the regression in the exposed group is “steeper,” that is, the vital capacity decays significantly faster with age. This corresponds to a smaller slope parameter for the exposed group since in this case the slopes are negative (Figure 14.5). The inference is supported by the following MATLAB code.

```matlab
xlsread vitalcapacity.xlsx;
twos = ans;
x1 = twos( twos(:,3) > 0, 1); y1 = twos( twos(:,3) > 0, 2);
x2 = twos( twos(:,3) ==0, 1); y2 = twos( twos(:,3) ==0, 2);
n1 = length(x1); n2 = length(x2);
SXX1 = sum((x1 - mean(x1)).^2); %4.3974e+003
```
14.5 Testing the Equality of Two Slopes*  703

\[
SXX2 = \text{sum}((x2 - \text{mean}(x2)).^2) \%6.1972e+003 \\
SYY1 = \text{sum}((y1 - \text{mean}(y1)).^2) \%26.5812 \\
SYY2 = \text{sum}((y2 - \text{mean}(y2)).^2) \%20.6067 \\
SXY1 = \text{sum}((x1 - \text{mean}(x1))\cdot(y1 - \text{mean}(y1))) \%-236.3850 \\
SXY2 = \text{sum}((x2 - \text{mean}(x2))\cdot(y2 - \text{mean}(y2))) \%-189.7116 \\
b1_1 = \frac{SXY1}{SXX1} \%-0.0538 \\
b1_2 = \frac{SXY2}{SXX2} \%-0.0306 \\
SSE1 = SYY1 - (SXY1)^2/SXX1 \%13.8741 \\
SSE2 = SYY2 - (SXY2)^2/SXX2 \%14.7991 \\
s2 = \frac{(SSE1 + SSE2)/(n1 + n2 - 4)}{\%0.3584} \\
s = \sqrt{s2} \%0.5987 \\
sbeb2 = s \times \sqrt{\frac{1}{SXX1} + \frac{1}{SXX2}} \%0.0118 \\
t = \frac{(b1_1 - b1_2)}{seb1b2} \%-1.9606 \\
pval = \text{tcdf}(t, n1 + n2 - 4) \%0.0267
\]

\[\text{Fig. 14.5} \text{ Samples from exposed (red) and unexposed (green) groups with fitted regression lines. The slopes of the two regressions are significantly different with a } p\text{-value smaller than 3\%.} \]

Thus, the hypothesis of equality of slopes is rejected with a \( p\)-value of 2.67\%.

**Remark.** Since the distribution of \( t\)-statistic is calculated under \( H_0\), which assumes parallel regression lines, the more natural estimator \( s22\), in place of \( s2\), takes into account this fact. The number of degrees of freedom in the \( t\)-statistic changes to \( n1 + n2 - 3\). The changes in the inference are minimal, as evidenced from the MATLAB code accounting for \( s22\):

\[
\%s22 \text{ accounts for equality of slopes:} \\
s22 = (SYY1 + SYY2 - \ldots) \\
    (SXY1 + SXY2)^2/(SXX1 + SXX2))/(n1 + n2 - 3) \%0.3710 \\
s = \sqrt{s22} \%0.6091 \\
seb1b2 = s \times \sqrt{\frac{1}{SXX1} + \frac{1}{SXX2}} \%0.0120
\]
For the case of the confidence interval, estimator $s^2$ and $n_1 + n_2 - 4$ degrees of freedom should be used.

### 14.6 Multiple Regression

It is often the case that in an experiment leading to regression analysis more than a single covariate is available. For example, in Chapter 2, page 32, we discussed an experiment in which two indexes of the amount of body fat (Siri and Brozek indexes) were calculated from the body density measure. In addition, a variety of body measurements, including weight, height, adiposity, neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrist, were recorded. Recall that the body density measure is complicated and potentially unpleasant, since it is taken by submerging the subject in water. Therefore, it is of interest to ask whether the Brozek index can be well predicted using the nonintrusive measurements.

If $x_1, x_2, \ldots, x_k$ are variables, covariates, or predictors, and we have $n$ joint measurements of covariates and the response, $x_{i1}, x_{i2}, \ldots, x_{ik}, y_i$, $i = 1, 2, \ldots, n$, then multiple regression expresses the response as a linear combination of covariates, plus an intercept and an additive error:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, i = 1, \ldots, n.$$  

The errors $\epsilon_i$ are assumed to be independent and normal with mean 0 and constant variance $\sigma^2$. We will denote by $k$ the number of covariates, but the number of parameters in the model is $p = k + 1$ because the intercept $\beta_0$ should be added. To avoid confusion, we will mostly use $p$ to represent the number of parameters in all expressions that involve dimensions and derived statistics.

As in the case of a single predictor, we will be interested in estimating and testing the coefficients, error variance, and mean and prediction responses. However, multiple regression brings several new challenges when compared to a simple regression. The two main challenges are (i) the possible presence of multicollinearity among the covariates, that is, covariates being correlated among themselves, and (ii) a multitude of possible models and the need to find the best model by identifying the “best” subset of predictors. A synonym for multiple regression is multivariable regression. Sometimes the multiple regression is wrongly termed multivariate regression; this terminology is reserved for the case where the response $y$ is multivariate, which is not the case here.
14.6.1 Matrix Notation

The regression equations for all \( n \) observations \((x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), i = 1, \ldots, n\), can be written as

\[
y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \epsilon_1,
\]

\[
y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \epsilon_2,
\]

\[\vdots\]

\[
y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \epsilon_n,
\]

and also in convenient matrix form as

\[
y = X\beta + \epsilon,
\]

where

\[
y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},
\]

\[
X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},
\]

\[
\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},
\]

\[
\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}
\]

Note that \( y \) and \( \epsilon \) are \( n \times 1 \) vectors, \( X \) is an \( n \times p \) matrix, and \( \beta \) is a \( p \times 1 \) vector. Here \( p = k + 1 \). To find the least-squares estimator of \( \beta \), one minimizes the sum of squares:

\[
\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2.
\]

The minimizing solution

\[
b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}
\]

satisfies the system (normal equations)
\[ X'Xb = X'y, \]  
(14.3)

and the least-squares estimator of \( \beta \) is

\[ b = (X'X)^{-1}X'y. \]

The fitted values are obtained as

\[ \hat{y} = Xb = X(X'X)^{-1}X'y, \]

and the residuals are

\[ e = y - \hat{y} = y - Xb = y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y, \]

where \( I \) is an \( n \times n \) identity matrix.

The matrix \( H = X(X'X)^{-1}X' \) that appears in the expressions for fitted values and residuals is important in this context; it is called the *hat* matrix. In terms of the hat matrix \( H \),

\[ \hat{y} = Hy, \quad \text{and} \quad e = (I - H)y. \]

Matrices \( H \) and \( I - H \) are projection matrices, and the \( n \)-dimensional vector \( y \) is projected to \( \hat{y} \) by \( H \) and to the residual vector \( e \) by \( I - H \). Any projection matrix \( A \) is *idempotent*, which means that \( A^2 = A \). Simply put, a projection of a projection will be the same as the original projection. Geometrically, vectors \( \hat{y} \) and \( e \) are orthogonal since the product of their projection matrices is 0. Indeed, because \( H \) is idempotent, \( H(I - H) = H - H^2 = H - H = 0 \).

The errors \( \epsilon_i \) are independent and the variance of vector \( e \) is \( \sigma^2I \).

We will illustrate some concepts from multiple regression using dataset \( \text{fat.dat} \); all calculations are part of MATLAB script \( \text{fatregdiag.m} \):

```matlab
load 'fat.dat'
casen = fat(:,1); % case number
broz = fat(:,2); % dependent variable
siri = fat(:,3); % function of dens
densi = fat(:,4); % an intrusive measure
% below are the predictors (except ffwei)
age = fat(:,5); weight = fat(:,6); height = fat(:,7);
adiposi = fat(:,8); % adiposity is BMI index=weight/height^2
ffwei = fat(:,9); % fat free weight, excluded from predictors
% since it involves body fat and brozek
neck = fat(:,10); chest = fat(:,11); abdomen = fat(:,12);
hip = fat(:,13); thigh = fat(:,14); knee = fat(:,15);
ankle = fat(:,16); biceps = fat(:,17); forearm = fat(:,18);
wrist = fat(:,19);
vecones = ones(size(broz)); % necessary for the intercept

disp('=======================================================')
```
14.6 Multiple Regression

```
disp(' p = 15, 14 variables + intercept')
disp('=======================================================')
Z = [age weight height adiposi neck chest abdomen ...
    hip thigh knee ankle biceps forearm wrist];
X = [vecones Z];
Y = broz
% X is design matrix, n x p where n is the number of subjects
% and p is the number of parameters, or number of predictors+1.
% varnames = ['intercept=0' 'age=1' 'weight=2' 'height=3'
% 'adiposi=4' 'neck=5' 'chest=6' 'abdomen=7' 'hip=8' 'thigh=9'
% 'knee=10' 'ankle=11' 'biceps=12' 'forearm=13' 'wrist=14'];
[n, p] = size(X)
b = inv(X' * X) * X' * Y;
H = X * inv(X' * X) * X';
max(max(H * H - H)); % 0 since H is projection matrix
Yhat = H * Y; % or Yhat = X * b;

14.6.2 Sums of Squares and an ANOVA Table

Sums of squares, SST, SSR, and SSE, for multiple regression have simple expressions in matrix notation. Here we introduce matrix \( J \), which is an \( n \times n \) matrix in which each element is 1. The total sum of squares can be calculated as

\[
SST = y' y - \frac{1}{n} y' J y = y' \left( I - \frac{1}{n} J \right) y.
\]

The error sum of squares is

\[
SSE = e' e = y' (I - H)' (I - H) y = y' (I - H) y
\]

because \( I - H \) is a symmetric projection matrix.

By taking the difference, we obtain

\[
SSR = SST - SSE = y' \left( H - \frac{1}{n} J \right) y.
\]

The number of degrees of freedom for SST, SSR, and SSE are \( n - 1 \), \( p - 1 \), and \( n - p \), respectively. Thus, the multiple regression ANOVA table is as follows:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>( p - 1 )</td>
<td>SSR</td>
<td>MSR</td>
<td>( \frac{SSR}{p-1} )</td>
<td>( F = \frac{MSR}{MSE} ) ( P(F_{p-1,n-p} &gt; F) )</td>
</tr>
<tr>
<td>Error</td>
<td>( n-p )</td>
<td>SSE</td>
<td>MSE</td>
<td>( \frac{SSE}{n-p} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n-1 )</td>
<td>SST</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
where large values of $F$ are critical for $H_0$, which states that the covariates $x_1, \ldots, x_k$ do not influence the response $y$. Formally, the null hypothesis is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

while the alternative is that at least one $\beta_i, i = 1, \ldots, k$ is not 0.

As in the simple regression, $R^2$ is the coefficient of determination,

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Adding more variables to a regression always increases $R^2$, even when the added covariates have nothing to do with the experiment. If two models have comparable $R^2$s, then, according to Ockham’s razor,\(^3\) the simpler model should be preferred, and adding new variables to the regression should be penalized. One way to achieve this is via an adjusted coefficient of determination,

$$R^2_{adj} = 1 - \frac{n - 1}{n - p} \frac{SSE}{SST},$$

which is one of the criteria for comparing models.

The estimator of the error variance $\sigma^2$ is $MSE$. We can find the confidence intervals of the components of $e$ and $\hat{y}$ using, respectively, the diagonal elements of covariance matrices $MSE \times (I - H)$ and $MSE \times H$.

\(^3\) *Pluralitas non est ponenda sine necessitate*, which translates into English as “Plurality should not be posited without necessity” (William of Ockham, 1287–1347).
14.6.3 Inference About Regression Parameters and Responses

The covariance matrix for a vector of estimators of regression coefficients $\mathbf{b}$ is equal to

$$s_b^2 = \text{MSE} \times (\mathbf{X}'\mathbf{X})^{-1}.$$ 

Its $(j + 1)$st diagonal element is an estimator of variance for $b_j$, and off-diagonal elements at the position $(j + 1, k + 1)$ are estimators of covariances between $b_j$ and $b_k$, where $j,k = 0,...,p-1$. Note that here we count the rows/columns of $p \times p$ matrix $(\mathbf{X}'\mathbf{X})^{-1}$ from 1 to $p$, while the corresponding indices in $\mathbf{b}$ run from 0 to $p - 1$.

When finding a confidence interval or testing a hypothesis about a particular $\beta_j$, $j = 0,...,p-1$, we use $b_j$ and $s_{b_j}$ (square root of $(j + 1)$st diagonal element of $s_b^2$) in the same way as in the univariate regression, only this time the test statistic $t$ has $n - p$ degrees of freedom instead of $n - 2$. Several subsequent MATLAB scripts are excerpts from the file `fatregdiag.m`:

```matlab
sig2 = MSE * inv(X' * X); % covariances among b's
sb = sqrt(diag(sig2));
tstats = b ./ sb;
pvals = 2 * tcdf(-abs(tstats), n-p);
disp('--------------------------------------
 disp(' var#   t     pval 
 disp('--------------------------------------
 [ (0:p-1)' tstats   pvals ]
%--------------------------------------
% var#   t     pval 
%--------------------------------------
% 0    -0.9430 0.3467
% 1.0000 1.8942 0.0594
% 2.0000 -1.6298 0.1045
% ... 
% 12.0000 0.9280 0.3543
% 13.0000 2.3243 0.0210
% 14.0000 -2.9784 0.0032

Note that in the fat example, the intercept is not significant ($p = 0.3467$), nor is the coefficient for variable #12 (biceps) ($p = 0.3543$), while the coefficient for variable #13 (forearm) is significant ($p = 0.0210$).

The regression response $y_m = \mathbf{x}^* \mathbf{b}$, evaluated at $\mathbf{x}^* = (1 \ x_1^* \ x_2^* \ ... \ x_k^*)$, has sample variance

$$s_{y_m}^2 = (\mathbf{x}^*) s_b^2 (\mathbf{x}^*)'.$$

For a prediction, the variance is, as in univariate regression, $s_{y_p}^2 = s_{y_m}^2 + \text{MSE}$. For the inference about regression response, the $t$-statistic with $n-p$
degrees of freedom is used. As an example, assume that for a “new” person with covariates \( x_h = (1 38 191 72 26 41 104 95 101.5 66 39 24 31 30 18.5) \), a prediction of the Brozek index is needed. The model gives a prediction of 19.5143, and the variances for mean response and individual response are estimated below. This is sufficient to calculate confidence intervals for the mean and individual responses.

\[
\text{Xh} = \begin{bmatrix} 1 & 38 & 191 & 72 & 26 & 41 & 104 & 95 & 101.5 & 66 & 39 & 24 & 31 & 30 & 18.5 \end{bmatrix};
\]
\[
\text{Yh} = \text{Xh} \ast \text{b} \quad \text{19.5143}
\]
\[
\text{sig2h} = \text{MSE} \ast \text{Xh} \ast \text{inv(X' * X)} \ast \text{Xh'};
\]
\[
\text{sig2hpre} = \text{MSE} \ast (1 + \text{Xh} \ast \text{inv(X' * X)} \ast \text{Xh'});
\]
\[
\text{sigh} = \sqrt{\text{sig2h}};
\]
\[
\text{sighpre} = \sqrt{\text{sig2hpre}};
\]
\[
\text{95\% CI's on the mean and individual responses}
\]
\[
[\text{Yh} - \text{tinv}(0.975, n-p) \ast \text{sigh}, \text{Yh} + \text{tinv}(0.975, n-p) \ast \text{sigh}]
\]
\[
[17.4347, 21.5940]
\]
\[
[\text{Yh} - \text{tinv}(0.975, n-p) \ast \text{sighpre}, \text{Yh} + \text{tinv}(0.975, n-p) \ast \text{sighpre}]
\]
\[
[11.3721, 27.6566]
\]

A Noninformative Bayesian Approach. A Bayesian inference in multiple linear regression is based on a prior on \((\beta, \sigma^2)\) and the likelihood
\[
f(y|\beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}
\]
\[
= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2} \left[ (y - \hat{y})'(y - \hat{y}) + (\beta - b)'(X'X)(\beta - b) \right] \right\},
\]
where \( b = (X'X)^{-1}X'y \) is the least squares estimator of \( \beta \), and \( \hat{y} = Xb \).

With noninformative prior
\[
\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2},
\]
the posterior for \((n - p)\frac{s^2}{\sigma^2} \) is \( \chi^2_{n-p} \), while the posterior for \( \beta \) is multivariate normal \( MVN_p(b, \sigma^2(X'X)^{-1}) \). The marginal posterior for \( \beta \), when \( \sigma^2 \) is integrated out, is multivariate t with location \( b \). Thus, the Bayes estimator for \( \beta \) coincides with \( b \), which is also traditional estimator (MLE). The Bayes estimator of \( \sigma^2 \) is \( \frac{n-p}{n-p-2}s^2 \), for \( s^2 = \frac{1}{n-p}(y - \hat{y})'(y - \hat{y}) \), and \( n > p + 2 \).

The \((1 - \alpha)100\%\) HPD set for \( \beta \) is
\[
\left\{ \beta \left| (\beta - b)'(X'X)(\beta - b) \leq p s^2 F_{1-s,p,n-p} \right. \right\},
\]
where $F_{1 - \alpha, p, n - p}$ is the $(1 - \alpha)$ quantile of an $F_{p, n - p}$-distribution.

For any particular component $\beta_j$,

$$t = \frac{\hat{\beta}_j - b_j}{s\sqrt{c_{jj}}}$$

has $t_{n - p}$ posterior distribution. Here $c_{jj}$ is the $(j + 1)$th diagonal element of $(XX)^{-1}$.

A more general prior on $(\beta, \sigma^2)$ is multivariate normal-inverse gamma ($\mathcal{NIG}$). This conjugate prior is capable of incorporating various prior information about the parameters and the prior $\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$ can be obtained as a limiting case of a $\mathcal{NIG}$ prior. Chapter 9 of O’Hagan (1994) provides an excellent description of the resulting Bayesian inference.

### 14.7 Diagnostics in Multiple Regression

Three important deficiencies in a multiple linear model can be diagnosed: (i) the presence of outliers, (ii) the nonconstant error variance, and (iii) a possible suboptimal model selection. Although the exposition level of these diagnostic methods exceeds the level in introductory coverage of regression, multiple regression modeling is important in practice and provides an important step to understanding more sophisticated nonlinear models, such as generalized linear models. For this reason, we provide a basic overview of residual and influence analysis, as well as an assessment of multicollinearity and choice of model. For readers interested in a more comprehensive treatment of multivariable linear models, the book by Rawlings et al. (2001) is a comprehensive resource.

#### 14.7.1 Residual Analysis and Influence

In setting the regression model, we made several assumptions about population errors (independence, zero-mean, constant variance, normality). The residuals, defined as $e_i = y_i - \hat{y}_i$, can be thought of as observed errors if the model is correct and should confirm our assumptions. For this reason, the residuals are examined graphically (histograms, plots against fits $\hat{y}_i$, against particular predictors, or, when it makes sense, against their order). When individual data points fail the residual check, we may suspect outliers in an otherwise correct model. However, if the residual analysis shows systematic deviations (trends, nonconstant variance), the model should be questioned. An interesting example of importance of residual analysis was constructed by Anscombe (1973), see Exercise 14.11.
Next we discuss the ordinary residuals and three modifications more appropriate for the statistical analysis.

The ordinary residuals $e_i = y_i - \hat{y}_i$ are components of $(I - H)y$. The leverages $h_{ii}$ are diagonal elements of the hat matrix $H$. These are important descriptors of design matrix $X$ and explain how far $x_i$ is from $\bar{x}$. All leverages are bounded $1/n \leq h_{ii} \leq 1$, and their sum is $\sum h_{ii} = p$, the number of regression parameters. Although the errors in the regression are independent and with the same variance, the residuals are correlated and with different sample variances $s^2(1 - h_{ii})$, where $s^2 = \text{MSE}$.

The studentized residual is the ordinary residual divided by its standard deviation

$$r_i = \frac{e_i}{s \sqrt{1 - h_{ii}}}$$

and recalls the $t$-statistic. Such residuals are scale-free comparable, and values outside the interval $[-2.5, 2.5]$ are potential outliers. Sometimes these residuals are called internally studentized, since the standard deviation $s = \sqrt{\text{MSE}}$ depends on the $i$th observation.

Externally studentized residuals, also called R-Student residuals, are measures of influence of the $i$th observation $(y_i, x_i)$ on the $i$th residual. Instead of $s$, the residuals are studentized by an external standard deviation,

$$s_{-i} = \sqrt{\frac{(n - p)s^2 - e_i^2 / (1 - h_{ii})}{n - p - 1}}. \quad (14.4)$$

This external estimate of $\sigma$ comes from the model fitted without the $i$th observation; however, the refitting is not necessary due to a simple expression in (14.4). Externally studentized residuals

$$t_i = \frac{e_i}{s_{-i} \sqrt{1 - h_{ii}}}$$

can be tested, since they are $t$ distributed with $n - p - 1$ degrees of freedom. Of course, if multiple residuals are tested simultaneously, then it should be done in the spirit of multiple hypothesis testing (Section 9.9).

PRESS (acronym for prediction sum of squares; Allen, 1974) residuals $e_{i,-i} = e_i / (1 - h_{ii})$ also remove the impact of the $i$th observation $(y_i, x_i)$ on the fit at $x_i$. This is a cross-validatory residual that measures how a model built without using the $i$th observation would predict the $i$th response.

For model assessment, the statistic PRESS is useful. It is defined as a sum of squares of PRESS residuals,

$$\text{PRESS} = \sum e_{i,-i}^2$$
and used in defining the *prediction* $R^2$,

$$R^2_{\text{pred}} = 1 - \frac{\text{PRESS}}{\text{SST}}.$$  

Here $\text{SST}$ is the total sum of squares $\sum (y_i - \overline{y})^2$. The ordinary $R^2$ is defined as $1 - \text{SSE}/\text{SST}$, and in $R^2_{\text{pred}}$ the $\text{SSE}$ is replaced by $\text{PRESS}$. Since the *average prediction error* is defined as $\sqrt{\text{PRESS}/n}$, good models should have a small PRESS.

The following table summarizes standard types of residuals used in residual analysis:

<table>
<thead>
<tr>
<th>Type</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ordinary residuals</td>
<td>$e_i = y_i - \hat{y}_i$</td>
</tr>
<tr>
<td>2. Studentized residuals</td>
<td>$r_i = \frac{e_i}{s \sqrt{1-h_{ii}}}$</td>
</tr>
<tr>
<td>3. Externally studentized residuals</td>
<td>$t_i = \frac{e_i}{s \sqrt{1-h_{ii}}}$</td>
</tr>
<tr>
<td>4. Prediction sum of squares residuals (PRESS)</td>
<td>$e_i_{-i} = \frac{e_i}{1-h_{ii}}$</td>
</tr>
</tbody>
</table>

$\text{DFBETAS}$ stands for *difference in betas*. It measures the influence of the $i$th observation on $\beta_j$:

$$\text{DFBETAS}_{ij} = \frac{b_j - b_j(-i)}{s_i \sqrt{c_{jj}}}$$

where $b_j$ is the estimator of $\beta_j$, $b_j(-i)$ is the estimator of $\beta_j$ when the $i$th observation is excluded, and $c_{jj}$ is the $(j+1)$st diagonal element in $(X'X)^{-1}$. Large $\text{DFBETAS}$ may indicate which predictor might be influential. The recommended threshold is $2/\sqrt{n}$. The black boxes in Figure 14.8 are at combinations ($\beta_0 - \beta_{14}$ vs. indices of observations) for which $\text{abs}(\text{DFBetas}>2/\sqrt{n})$.

Since several $\text{DFBETAS}$ can be large, it is useful to pay attention to those corresponding to large $\text{DFFITS}$. The $\text{DFFITS}$ measure the influence of the $i$th observation on the prediction $\hat{y}_i$:

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{(-i)} \sqrt{h_{ii}}} = \sqrt{\frac{h_{ii}}{1-h_{ii}}} \cdot \frac{e_i}{s_{-i} \sqrt{1-h_{ii}}}.$$  

The value $\hat{y}_{i,-i}$ is the prediction of $y_i$ on the basis of a regression fit without the $i$th observation. The observation is considered influential if its $\text{DFFITS}$ value exceeds $2\sqrt{p/n}$.

A measure related computationally to $\text{DFFITS}$ is Cook’s distance, $D_i$:
\[ D_i = (DFFITS_i)^2 \cdot \frac{s^2_i}{p s^2} \]

Cook’s distance measures the effect of the \( i \)th observation on the whole vector \( \mathbf{b} = \hat{\mathbf{b}} \). An observation is deemed influential if its Cook’s distance exceeds \( 4/n \).

Figure 14.7 shows ordinary residuals plotted against predicted values \( \hat{y} \). The radii of circles are proportional to \( |Dffits| \) in panel (a) and to Cook’s distance in panel (b).

Influential observations are not necessarily outliers and should not be eliminated from a model only on the basis of their influence. Such observations can be identified by their influence on a predicted value. One often finds predictions of the \( i \)th response \( \hat{y}_{i,-i} \) in a regression in which the \( i \)th case \((y_i, x_i)\) is omitted (Figure 14.6).

```matlab
%prediction of \( y_i \) with \( i \)th observation removed
%hat \( y_i \)(-i)
ind = 1:n;
Yhati = [];
for i = 1:n
    indi = find(ind ~= i);
    Yi = Y(indi);
    Xi=X(indi,:);
    bi = inv(Xi’ * Xi) * Xi'* Yi;
    Yhatii = X(i,:) * bi;
    Yhati =[Yhati; Yhatii];
end
Yhati %prediction of \( y_i \) without \( i \)th observation

%-------------------------------------------------
```

Fig. 14.6 Predicted responses \( \hat{y}_i \) (blue) and predicted responses \( \hat{y}_{i,-i} \) (red). Large changes in prediction signify an influential observation.
14.7 Diagnostics in Multiple Regression

%============== residual analysis==================

hii = diag(H); % leverages
resid = (I - H)*Y; % ordinary residuals
sresid = sqrt(MSE .* (1-hii));
stresid = resid./sresid % studentized residuals
%---------studentized deleted residuals---
di = Y - Yhati; % or di = resid./(1-hii)
% di is also called PRESS residual
PRESS = sum(di.^2)
R2pred = 1-PRESS/SST % R^2 predictive

sminusi = sqrt(((n-p)*MSE*ones(n,1) - resid.^2./(1-hii))/(n-p-1)); % stdev(-i)
ti = resid ./ (sminusi .* sqrt(1-hii)) % externally studentized residuals
% outliers based on leverage = hii
outli=hii/mean(hii);
find(outli > 3)
% 31 36 39 41 42 86 159 175 206

% influential observations
Dffits = ti .* sqrt( hii ./ (1-hii)) % influence ith to ith
find(abs(Dffits) > 2 * sqrt(p/n));
% 31 39 42 86 128 140 175 207 216 221 250

CooksD = resid.^2 .* (hii./(1-hii).^2)/(p * MSE) % influence if ith to all
find(CooksD > 4/n) % find influential
% 31 39 42 86 128 175 207 216 221 250

Fig. 14.7 Ordinary residuals plotted against predicted values \( \hat{y} \). The radii of circles are proportional to |Dffits| in panel (a) and to Cook’s distance in panel (b). The observations with the largest circles are in both cases the 42nd and the 39th.

% DFBETAS - influence if ith obs on jth coefficient
Regression

cii = diag(inv(X' * X));
DFBetas =[];
for i = 1:n
  indi = find(ind ~= i);
  Yi = Y(indi);
  Xi=X(indi,:);
  bi = inv(Xi' * Xi) * Xi'* Yi;
  Hi = Xi * inv(Xi' * Xi) * Xi';
  SSEi = Yi' * (eye(n-1) - Hi) * Yi;
  MSEi = SSEi./(n-p-1);
  DFBetasi = (b - bi)./sqrt(MSEi .* cii) ;
  DFBetas = [DFBetas; DFBetasi'];
end

Fig. 14.8 DFBETAS: The x-axis enumerates $\beta_0 - \beta_{14}$, while on the y-axis are plotted the indices of observations. The black boxes are at combinations for which $\text{abs}(DFBetas)>2/sqrt(n))$.

14.7.2 Multicollinearity

The multicollinearity problem in regression concerns the correlation among the predictors. Suppose that matrix $X$ contains two collinear columns $x_i$ and $x_j = 2x_i$. Obviously covariates $x_i$ and $x_j$ are linearly dependent, and $x_j$ does not bring any new information about the response. This collinearity makes matrix $X$ not of full rank and $X'X$ singular, that is, not invertible, and normal equations (14.3) have no unique solution. In reality, if multicollinearity is present, then matrix $X'X$ is not singular, but near-singular, in the sense that its determinant is close to 0, making the inversion of $X'X$, and consequently the solution $\hat{\beta}$, numerically unstable. This happens when either two or more variables are highly correlated or when a variable has a small variance (and, in a sense, is correlated to the intercept).

In addition to numerical instability in computing the least squares estimators, the multicollinearity can lead to inferential inconsistencies. Mul-
Diagnostics in Multiple Regression

Multicollinear predictors tend to have $t$-statistics for their corresponding coefficients shrunk toward 0, leading to the possibility of having a significant regression, as assessed by statistic $F$, in which none of the individual predictors turns out significant.

There are many measures of multicollinearity discussed in the literature. We will discuss the condition number, condition indexes, and multicollinearity index, which are global measures, and the variance inflation factor, which is linked to a particular predictor.

Let $Z_{n \times k}$ be a matrix obtained from the design matrix $X$ by omitting the vector of ones, $Z = X(:,2:end)$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ be ordered eigenvalues of correlation matrix of the predictors, in MATLAB, $corr(Z)$. The condition number (Belsley et al., 1980) is defined as the square root of the ratio of the largest and smallest eigenvalues,

$$K = \sqrt{\frac{\lambda_1}{\lambda_k}}.$$ 

Values for $K$ starting at around 10 are concerning, values between 30 and 100 influence the results, and values over 100 indicate a serious collinearity problem.

The condition indexes for $corr(Z)$ defined as

$$K_j = \sqrt{\frac{\lambda_1}{\lambda_j}}, \; j = 1,\ldots,k,$$

are indicative of possible presence of multiple near-linear relationships among the predictors. For example, if three condition indexes are large, say exceed 30, this would mean that among predictors there are three competing near-linear dependencies independent of each other.

The multicollinearity index ($MCI$) suggested by Thisted (1980) is defined as

$$MCI = \sum_{j=1}^{k} \left( \frac{\lambda_k}{\lambda_j} \right)^2.$$ 

Here, $\lambda_k$ is the smallest eigenvalue of $corr(Z)$. The $MCI$ falls between 1 and $k$, it is equal to 1 for exactly collinear variables, and to $k$ for orthogonal variables. The values of $MCI$ close to 1 indicate high collinearity, values larger than 2 are not alarming.

The variance inflation factor ($VIF$) explains the extent of correlation of a particular variable $x_j$ to the rest of predictors. It is defined as
\( VIF_j = \frac{1}{1 - R_j^2} \)

where \( R_j^2 \) is the coefficient of determination in regression of \( x_j \) to the rest of predictors. The name variance inflation factor comes from an alternate expression for variance of \( b_j \),

\[
\text{Var}(b_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}, \quad j = 1, \ldots, p - 1,
\]

where \( x_{ij} \) are elements of design matrix \( X \) without the column of 1’s corresponding to the intercept. It is obvious that when \( R_j \) approaches 1, the variance of \( b_j \) explodes. \( VIFs \) exceeding 10 are considered serious.

Computation of \( VIFs \)'s is simple. One finds the inverse of the correlation matrix for the predictors, The diagonal elements of this inverse are the \( VIFs \),

\[
\text{diag}(\text{inv}(\text{corr}(X(:,2:end))))),
\]

where \( X \) is the design matrix.

As a global measure of multicollinearity an average \( VIF \) is used,

\[
\overline{VIF} = \frac{1}{k} \sum_{j=1}^{k} VIF_j, \quad k = p - 1.
\]

Values of \( \overline{VIF} \) substantially larger than 1 indicate a multicollinearity problem.

A note of caution: \( VIF \) diagnostic sometimes can miss a problem since the intercept is not included in the analysis.

Multicollinearity can be diminished by excluding problematic variables causing the collinearity in the first place. Alternatively, groups of variables can be combined and merged into a single variable.

Another way to diminish multicolinearity is to keep all variables but “condition” matrix \( X'X \) by adding \( kI \), for some \( k > 0 \), to the normal equations. This is known as ridge regression (Hoerl and Kennard, 1970). There is a tradeoff: the solutions of \( (X'X + kI)\hat{\beta} = X'y \) are more stable, but some bias is introduced.

```matlab
% Z = X(:,2:end);
RXX = corr(Z);
lambda = eig(RXX);
K = sqrt(max(lambda)/min(lambda)) \% 19.3713
K1 = sqrt(max(lambda)/lambda);
K1'\%
\% 1.0000 2.4075 2.9039 3.6422 3.8436 5.2447 5.4643
MCI = sum((min(lambda)/lambda).^2) \% 1.5853
VIF = diag(inv(RXX));
VIF'\%
```
Alternatively, for most of the above calculations one can use MATLAB’s 
`regstats` or `diagnostics.m`:

```matlab
s = regstats(Y,Z,'linear','all');
[index,res,stud_res,lev,DFFITS1,Cooks_D,DFBETAS]=diagnostics(Y,Z);
```

### 14.7.3 Variable Selection in Regression

Model selection involves finding a subset of predictors from a large number of potential predictors that is optimal in some sense.

We defined the coefficient of determination $R^2$ as the proportion of model-explained variability, and it seems natural to choose a model that maximizes $R^2$. It turns out that this is not a good idea since the maximum will always be achieved by that model that has the maximal number of parameters. It is a fact that $R^2$ increases when even a random or unrelated predictor is included.

The adjusted $R^2$ penalizes the inclusion of new variables and represents a better criterion for choosing a model. However, with $p$ parameters there would be $2^p$ possible candidate models, and even for moderate $p$ this could be a formidable number.

There are two mirroring procedures, forward selection and backward selection, that are routinely employed in cases where checking all possible models is infeasible. Forward selection proceeds in the following way:

**STEP 1.** Start with the intercept-only model. Choose the predictor that has the largest $R^2$ among the models with a single variable. Call this variable $x_1$.

**STEP 2.** Assume that the model already has $m$ variables, $x_1, \ldots, x_m$, for some $m \geq 1$. Select the variable $x_{m+1}$ that gives the maximal increase to $R^2$ and refit the model.

**STEP 3.** Denote by $SSR(x_1, \ldots, x_m)$ the regression sum of squares for a regression fitted with variables $x_1, \ldots, x_m$. Then $R(x_{m+1} | x_1, \ldots, x_m) = SSR(x_1, \ldots, x_{m+1}) - SSR(x_1, \ldots, x_m)$ is the contribution of the $(m+1)$st variable and it is considered significant if

$$R(x_{m+1} | x_1, \ldots, x_m) / MSE(x_1, \ldots, x_{m+1}) > F_{1, n-m-1, 1-\alpha}$$

(14.5)

where $MSE(x_1, \ldots, x_{m+1})$ the mean square error for a regression fitted with variables $x_1, \ldots, x_{m+1}$. If relation (14.5) is satisfied, then variable $x_{m+1}$ is included in the model. Increase $m$ by one and go to **STEP 2**.

If relation (14.5) is not satisfied, then the contribution of $x_{m+1}$ is not significant, in which case go to **STEP 4**.

**STEP 4.** Stop with the model that has $m$ variables. END

The $MSE$ in (14.5) was estimated from the full model. Note that the forward selection algorithm is “greedy” and chooses the single best improving
variable at each step. This, of course, may not lead to the optimal model since in reality variable \( x_1 \), which is the best for one-variable models, may not be included in the best two-variable model.

Backward stepwise regression starts with the full model and removes variables with insignificant contributions to \( R^2 \). Seldom do these two approaches end with the same candidate model.

MATLAB’s Statistics Toolbox has two functions for stepwise regression: \texttt{stepwisefit}, a function that proceeds automatically from a specified initial model and entrance/exit tolerances, and \texttt{stepwise}, an interactive tool that allows you to explore individual steps in a process.

An additional criterion for the goodness of a model is the Mallows \( C_p \). This criterion evaluates a proposed model with \( k \) variables and \( p = k + 1 \) parameters. The Mallows \( C_p \) is calculated as

\[
C_p = (n - p) \frac{s^2}{\hat{\sigma}^2} - n + 2p,
\]

where \( s^2 \) is the MSE of the candidate model and \( \hat{\sigma}^2 \) is an estimator of \( \sigma^2 \), usually taken to be the best available estimate. The MSE of the full model is typically used as \( \hat{\sigma}^2 \).

A common misinterpretation is that in \( C_p \), \( p \) is referred to as the number of predictors instead of parameters. This is correct only for models without the intercept (or when 1 from the vector of ones in the design matrix is declared as a predictor).

Adequate models should have a small \( C_p \) that is close to \( p \). Typically, a plot of \( C_p \) against \( p \) for all models is made. The “southwesternmost” points close to the line \( C_p = p \) correspond to adequate models. The \( C_p \) criterion can also be employed in forward and backward variable selection as a stopping rule.

### 14.7.4 Bayesian Model Selection in Multiple Regression

Next, we revisit \texttt{fat.dat} with some Bayesian analyses. Four competing models are compared using the Laud–Ibrahim predictive criterion, LI. Models with smaller LI are favored.

Laud and Ibrahim (1995) argue that agreement of model-simulated predictions and original data should be used as a criterion for model selection. If for \( y_i \) responses \( \hat{y}_{i,new} \) are hypothetical replications according to the posterior predictive distribution of competing model parameters, then

\[
LI = \sum_{i=1}^{n} (E\hat{y}_{i,new} - y_i)^2 + \text{Var}(\hat{y}_{i,new})
\]
measures the discrepancy between the observed and model-predicted data. A smaller LI is better. The file `fat.odc` performs a Laud–Ibrahim Bayesian model selection and prefers model #2 of the four models analyzed.

```plaintext
#fat.odc
model{
  for(j in 1:N ){
    # four competing models
  }
  #LI - Laud-Ibrahim Predictive Criterion. LI-smaller-better
  for(i in 1:4 ){
    tau[i] ~ dgamma(2,32)
    LI[i] <- sqrt( sum(D2[i,]) + pow(sd(broz.new[i,]),2))
    # data sets 1-4 for different models
    for (j in 1:N) {
      broz2[i,j] <- broz[j]
      broz2[i,j] ~ dnorm(mu[i,j],tau[i])
      broz.new[i,j] ~ dnorm(mu[i,j],tau[i])
      D2[i,j] <- pow(broz[j]-broz.new[i,j],2)
    }
  }
  # Compare predictive criteria between models i and j
  # Comp[i,j] is 1 when LI[i]<LI[j], i-th model better.
  for (i in 1:3) { for (j in i+1:4 ) {
    Comp[i,j] <- step(LI[j]-LI[i])
  }
  # priors
  for (j in 1:15) { b1[j] ~ dnorm(0,0.001) }
  for(j in 1:4) { b2[j] ~ dnorm(0,0.001) }
  b4[j] ~ dnorm(0,0.001)
  for(j in 1:2) { b3[j] ~ dnorm(0,0.001) }
}

#DATA 1: Load this first
list(N = 252)

# DATA2: Then load the variables
broz[] age[] wei[] hei[] ... biceps[] forea[] wrist[]
12.6 23 154.25 67.75 ... 32.0 27.4 17.1
23.4 38.5 93.6 83.00 ... 30.5 28.9 18.2
...248 lines deleted...
25.3 72 198.75 70.50 ... 30.5 29.4 19.8
38.7 74 207.50 70.00 ... 33.7 30.0 28.9
END
```
The output is given in the table below. Note that even though the posterior mean of \text{LI}[4] is smaller that that of \text{LI}[2], it is the posterior median that matters. Model #2 is more frequently selected as the best compared to model #4.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>val2.5pc</th>
<th>median</th>
<th>val97.5pc</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>LI[1]</td>
<td>186.5</td>
<td>209.2</td>
<td>17.63</td>
<td>84.0</td>
<td>104.6</td>
<td>910.5</td>
<td>1001</td>
<td>20000</td>
</tr>
<tr>
<td>LI[2]</td>
<td>96.58</td>
<td>23.46</td>
<td>1.924</td>
<td>85.08</td>
<td>93.14</td>
<td>131.2</td>
<td>1001</td>
<td>20000</td>
</tr>
<tr>
<td>LI[3]</td>
<td>119.6</td>
<td>5.301</td>
<td>0.03587</td>
<td>109.4</td>
<td>119.5</td>
<td>130.2</td>
<td>1001</td>
<td>20000</td>
</tr>
<tr>
<td>LI[4]</td>
<td>94.3</td>
<td>4.221</td>
<td>0.0396</td>
<td>86.33</td>
<td>94.2</td>
<td>103.0</td>
<td>1001</td>
<td>20000</td>
</tr>
<tr>
<td>Comp[1,2]</td>
<td>0.2974</td>
<td>0.4571</td>
<td>0.02785</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1001</td>
<td>20000</td>
</tr>
<tr>
<td>Comp[1,3]</td>
<td>0.5844</td>
<td>0.4928</td>
<td>0.03939</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1001</td>
<td>20000</td>
</tr>
<tr>
<td>Comp[1,4]</td>
<td>0.3261</td>
<td>0.4688</td>
<td>0.03008</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1001</td>
<td>20000</td>
</tr>
<tr>
<td>Comp[2,3]</td>
<td>0.9725</td>
<td>0.1637</td>
<td>0.01344</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1001</td>
<td>20000</td>
</tr>
<tr>
<td>Comp[2,4]</td>
<td>0.5611</td>
<td>0.4963</td>
<td>0.01003</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1001</td>
<td>20000</td>
</tr>
<tr>
<td>Comp[3,4]</td>
<td>5.0E-5</td>
<td>0.007071</td>
<td>4.98E-5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1001</td>
<td>20000</td>
</tr>
</tbody>
</table>

More comprehensive treatment of Bayesian approaches in linear regression can be found in O’Hagan (1994) and Ntzoufras (2009).

### 14.8 Sample Size in Regression

The evaluation of power in a regression with \( p - 1 \) variables and \( p \) parameters (an intercept is present) requires specification of a significance level and precision. Suppose that we want a power such that a total sample size of \( n = 61 \) would make \( R^2 = 0.2 \) significant for \( \alpha = 0.05 \) and the number of predictor variables \( p - 1 = 3 \). Cohen’s effect size here is defined as \( f^2 = R^2/(1 - R^2) \). Unlike the ANOVA where the values of \( f^2 \approx 0.01 \) corresponded to small, \( f^2 \approx 0.0625 \) to medium, and \( f^2 \approx 0.16 \) to large effects, in regression the values of \( f^2 \approx 0.02 \) corresponded to small, \( f^2 \approx 0.15 \) to medium, and \( f^2 \approx 0.35 \) to large effects. Note that from \( f^2 = R^2/(1 - R^2) \) one gets \( R^2 = f^2/(1 + f^2) \), which can be used to check the adequacy of the elicited/required effect size.

The power, similar to ANOVA, is found using the noncentral \( F \)-distribution,

\[
1 - \beta = \mathbb{P}(F^{nc}(p - 1, n - p, \lambda) > F^{-1}(1 - \alpha, p - 1, n - p)),
\]

where \( \lambda = nf^2 \) is the noncentrality parameter.
Example 14.5. **Power Analysis in Regression.** For $p = 4$, $R^2 = 0.2$, that is, $f^2 = 0.25$ (a medium-to-large effect), and a sample size of $n = 61$, one gets $\lambda = 61 \times 0.25 = 15.25$, and a power of approximately 90%.

```matlab
p=4; n=61; lam=15.25;
1-nfcdf( finv(1-0.05, p-1, n-p), p-1, n-p, lam)
% ans = 0.9014
```

### 14.9 Linear Regression That Is Nonlinear in Predictors

In linear regression, “linear” concerns the parameters, not the predictors. For instance,

$$
\frac{1}{y_i} = \beta_0 + \frac{\beta_1}{x_i} + \epsilon_i, \quad i = 1, \ldots, n,
$$

and

$$
y_i = \epsilon_i \times \exp\{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\}, \quad i = 1, \ldots, n,
$$

are examples of a linear regression. There are many functions where $x$ and $y$ can be linearized by an obvious transformation of $x$ or $y$ or both; however, one needs to be mindful that in such transformations the normality and homoscedasticity of errors is often compromised. In such a case, fitting a regression is simply an optimization task without natural inferential support. Bayesian solutions involving MCMC are generally more informative regarding the inference (Bayes estimators, credible sets, predictions).

An example where errors are not affected by the transformation of variables is a polynomial relationship between $x$ and $y$ postulated as

$$
y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i, \quad i = 1, \ldots, n,
$$

which is in fact a linear regression. Simply, the $k$ predictors are $x_{1i} = x_i$, $x_{2i} = x_i^2$, $\ldots$, $x_{ki} = x_i^k$, and estimating the polynomial relationship is straightforward. The example below is a research problem in which a quadratic relationship is used.

**Example 14.6. Von Willebrand Factor.** Von Willebrand disease is a bleeding disorder caused by a defect or deficiency of a blood clotting protein called the von Willebrand factor. This glue-like protein, produced by the cells that line blood vessel walls, interacts with blood cells called platelets to form a plug that prevents bleeding. In order to understand the differential bonding mechanics underlying von Willebrand-type bleeding disorders, researchers at Georgia Tech studied the interactions between the
The mean stop time rolling parameter was calculated from frame-by-frame rolling velocity data collected at 250 frames per second. Mean stop time indicates the amount of time a cell spends stopped, so it is analogous to the bond lifetime. This parameter, being an indicator for how long a bond is stopped before the platelet moves again, can be used to assess the bond lifetime and off-rate (Yago et al., 2008).

For the purpose of exploring interactions between the force and the mean stop times, Ficoll 6% is added to increase the viscosity. Data are courtesy of Dr. Leslie Coburn. The mat file `coburn.mat` contains the structure `coburn` with data fields `coburn.fxssy`, where `x = 0, 6` is a code for Ficoll absence/presence and `y = 1, 2, 4, ..., 256` denotes the shear stress (in dyn/cm²). For example, `coburn.f0ss16` is a 243 × 1 vector of mean stop times obtained with no Ficoll, under a shear stress of 16 dyn/cm².

<table>
<thead>
<tr>
<th>Shear stress</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shear number</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Mean stop time</td>
<td>f0ss2 f0ss4 f0ss8 f0ss16 f0ss32 f0ss64 f0ss128 f0ss256</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample size</td>
<td>26</td>
<td>57</td>
<td>157</td>
<td>243</td>
<td>256</td>
<td>185</td>
<td>62</td>
<td>14</td>
</tr>
</tbody>
</table>

We fit a regression on the logarithm (base 2) of mean stop time \( \log_{2}\text{mst} \), with no Ficoll present, as a quadratic function of share stress number \( \log_{2}\text{(dyn/cm}^2\text{)} \). This regression is linear in parameters with two predictors, \( \text{shearn} \) and the squared shear, \( \text{shearn}^2 \).

The regression fit is

\[
\log_{2}\text{mst} = -6.2423 + 0.8532 \text{shear} - 0.0978 \text{shear}^2.
\]

The regression is significant (\( F = 63.9650, p = 0 \)); however, its predictive power is rather weak, with \( R^2 = 0.1137 \).

Figure 14.9 is plotted by the script `coburnreg.m`. %coburnreg.mat

```matlab
load 'coburn.mat';
mst=[coburn.f0ss2; coburn.f0ss4; coburn.f0ss8; coburn.f0ss16; coburn.f0ss32; coburn.f0ss64; coburn.f0ss128; coburn.f0ss256];
shearn = [1 * ones( 26,1); 2 * ones( 57,1); 3 * ones(157,1); 4 * ones(243,1); 5 * ones(256,1); 6 * ones(185,1); 7 * ones( 62,1); 8 * ones( 14,1)];
shearn2 = shearn.^2; %quadratic term

%design matrix
X = [ones(length(shearn),1) shearn shearn2];
[b,bint,res,resint,stats] = regress(log2(mst), X);```
14.10 Errors-in-Variables Linear Regression

Assume that in the context of regression both responses $Y$ and covariates $X$ are measured with error. This is a frequent scenario in research labs in which it would be inappropriate to apply standard linear regression, which assumes that covariates are designed and constant.

This scenario in which covariates are observed with error is called errors-in-variables (EIV) linear regression. There are several formulations for EIV regression (Fuller, 2006). For pairs from a bivariate normal distribution $(x_i, y_i), i = 1, \ldots, n$, the EIV regression model is

\[
y_i \sim \mathcal{N}(\beta_0 + \beta_1 \bar{\xi}_i, \sigma_y^2)
\]

\[
x_i \sim \mathcal{N}(\bar{\xi}_i, \sigma_x^2).
\]

In an equivalent form, the regression is $E[y_i] = \beta_0 + \beta_1 E[x_i] = \beta_0 + \beta_1 \bar{\xi}_i$, $i = 1, \ldots, n$, and the inference on parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ is made conditionally on $\bar{\xi}_i$. To make the model identifiable, parameter $\eta = \frac{\sigma_x^2}{\sigma_y^2}$ is assumed known. Note that we do not need to know individual variances, just their ratio.

The $R^2$ is about 11%, which is small, but represents improvement if only simple linear regression with covariate $\text{sharen}$ were used. Note that residuals are approximately normal (Figure 14.9b), as expected.

Fig. 14.9 (a) Quadratic regression on log mean stop time. (b) Residuals fitted with normal density.

\[
\text{b}_0 = -6.2423, \quad b_1 = 0.8532, \quad b_2 = -0.0978
\]

\[
\begin{array}{llll}
\text{R}^2 & F & p & \sigma_y^2 \\
0.1137 & 63.9650 & 0.0000 & 0.5856
\end{array}
\]
If the observations \((x_i, y_i), i = 1, \ldots, n\), produce sums of squares
\[ S_{xx} = \sum_i (x_i - \bar{x})^2, \quad S_{yy} = \sum_i (y_i - \bar{y})^2, \quad \text{and} \quad S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}), \]
then
\[
\hat{\beta}_1 = \frac{-(S_{xx} - \eta S_{yy}) + \sqrt{(S_{xx} - \eta S_{yy})^2 + 4\eta S_{xy}^2}}{2\eta S_{xy}}, \tag{14.7}
\]
\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.
\]
The estimators of the errors are
\[
\hat{\sigma}_x^2 = \frac{1}{2n} \times \frac{\eta}{1 + \eta \hat{\beta}_1^2} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2,
\]
\[
\hat{\sigma}_y^2 = \frac{\hat{\sigma}_x^2}{\eta}.
\]

When \(\eta = 1\), meaning variances of errors are the same, the solution in (14.7) coincides with a numerical least-squares minimization (so called orthogonal regression). However, for \(\eta = 0\), meaning no errors in covariates, \(\sigma_x^2 = 0\), we are back to the standard regression where \(\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}\) (Exercise 14.24).

**Example 14.7. Predicting Albumin from Plasma Volume.** Griffin et al. (1945) reported plasma volume (in cc) and circulating albumin (in g) for \(n = 58\) healthy males. Both quantities were measured with error, and it was assumed that the variance of plasma measurement exceeded the variance of circulating albumin by a factor of 200. Using EIV regression, establish an equation that would be used to predict circulating albumin from plasma volume. The data are given in \texttt{circalbumin.dat}, where the first column contains plasma volume and the second the measured albumin. The script \texttt{errorinvar.m} calculates the following EIV regression equation: \(\mathbb{E}y_i = 0.0521 \cdot \mathbb{E}x_i - 13.1619\). This straight line is plotted in red in Figure 14.10. For comparison, the standard regression is \(\mathbb{E}y_i = 0.0494x_i - 5.7871\), and it is plotted in black. Parameter \(\eta\) was set to 200, and the variances are estimated as \(\hat{s}_x^2 = 4852.8\) and \(\hat{s}_y^2 = 24.2641\).

### 14.11 Analysis of Covariance

Analysis of covariance (ANCOVA) is a linear model that includes two types of predictors: quantitative, like regression, and categorical, like ANOVA. It
can be formulated in quite general terms, but we will discuss the case of a single predictor of each kind.

The quantitative variable \( x \) is linearly connected with the response, and for a fixed categorical variable, the problem is exactly regression. However, for a fixed value of \( x \), the model is ANOVA with treatments/groups defined by the categorical variable.

The rationale behind the merging of the two models is that in a range of experiments, modeling the problem as regression only or as ANOVA only may be inadequate. By introducing a quantitative covariate to ANOVA, or equivalently groups/treatments to regression, we may better account for the variability in data and produce better modeling and prediction results.

We will analyze two illustrative examples. In the first example, which will be solved in MATLAB, the efficacies of two drugs for lowering blood pressure are compared by analyzing the drop in blood pressure after taking the drug. However, the initial blood pressure measured before the drug is taken should be taken into account since a drop of 50, for example, is not the same if the initial pressure was 90 as opposed to 180.

In the second example, which will be solved in WinBUGS, the measured response is the strength of synthetic fiber, and the two covariates are the fiber’s diameter (quantitative) and the machine on which the fiber was produced (categorical with three levels).

We assume the model

\[
y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \overline{x}) + \epsilon_{ij}, \ i = 1, \ldots, a; \ j = 1, \ldots, n,
\]

(14.8)

where \( a \) is the number of levels/treatments, \( n \) is a common sample size within each level, and \( \overline{x} = \frac{1}{an} \sum_{i,j} x_{ij} \) is the overall mean of \( xs \). The errors \( \epsilon_i \)
are assumed to be independent normal with mean 0 and constant variance \( \sigma^2 \).

For practical reasons the covariates \( x_{ij} \) are centered as \( x_{ij} - \bar{x} \) in order to simplify the expressions for the estimators. Let \( \bar{x}_i = \frac{1}{n} \sum x_{ij} \) be the \( i \)th treatment mean for the \( x \)s. The means \( \bar{y} \) and \( \bar{y}_i \) are defined analogously, \( \bar{y} = \frac{1}{an} \sum y_{ij} \) and \( \bar{y}_i = \frac{1}{n} \sum y_{ij} \).

In ANCOVA we calculate the sums of squares and mixed-product sums as

\[
S_{xx} = \sum_{i=1}^a \sum_{j=1}^n (x_{ij} - \bar{x})^2 \\
S_{xy} = \sum_{i=1}^a \sum_{j=1}^n (x_{ij} - \bar{x})(y_{ij} - \bar{y}) \\
S_{yy} = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y})^2 \\
T_{xx} = \sum_{i=1}^a (\bar{x}_i - \bar{x})^2 \\
T_{xy} = \sum_{i=1}^a (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) \\
T_{yy} = \sum_{i=1}^a (\bar{y}_i - \bar{y})^2 \\
Q_{xx} = \sum_{i=1}^a \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 = S_{xx} - T_{xx} \\
Q_{xy} = \sum_{i=1}^a \sum_{j=1}^n (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i) = S_{xy} - T_{xy} \\
Q_{yy} = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 = S_{yy} - T_{yy}
\]

We are interested in finding estimators for the parameters in model (14.8), the common mean \( \mu \), treatment effects \( \alpha_i \), regression slope \( \beta \), and the variance of the error \( \sigma^2 \).

The estimators are \( \hat{\mu} = \bar{y}, \hat{b} = \hat{\beta} = Q_{xy}/Q_{xx}, \) and \( \hat{\alpha}_i = \bar{y}_i - \bar{y} - b(\bar{x}_i - \bar{x}) \). The estimator of the variance, \( s^2 \), is \( s^2 = MSE = SSE/(a(n - 1) - 1) \), where \( SSE = Q_{yy} - Q_{xy}^2/Q_{xx} \).

If there are no treatment effects, that is, if all \( \alpha_i = 0 \), then the model is a plain regression and

\[
y_{ij} = \mu + \beta(x_{ij} - \bar{x}) + \epsilon_{ij}, \ i = 1, \ldots, a; \ j = 1, \ldots, n.
\]

In this reduced case the error sum of squares is \( SSE' = S_{yy} - S_{xy}^2/S_{xx} \), with \( an - 2 \) degrees of freedom. Thus, the test \( H_0 : \alpha_i = 0 \) is based on an \( F \)-statistic,

\[
F = \frac{(SSE' - SSE)/(a - 1)}{SSE/(a(n - 1) - 1)},
\]

that has an \( F \)-distribution with \( a - 1 \) and \( a(n - 1) - 1 \) degrees of freedom.

The test for regression \( H_0 : \beta = 0 \) is based on the statistic

\[
F = \frac{Q_{xy}^2/Q_{xx}}{SSE/(a(n - 1) - 1)}
\]
14.11 Analysis of Covariance

which has an $F$-distribution with 1 and $a(n - 1) - 1$ degrees of freedom.

Next we provide a MATLAB solution for a simple ANCOVA layout.

**Example 14.8. Kodlin’s Blood Pressure Experiment.** Kodlin (1951) reported an experiment that compared two substances for lowering blood pressure, denoted as substances A and B. Two groups of animals are randomized to the two substances and a decrease in pressure is recorded. The initial pressure is recorded. The data for the blood pressure experiment appear in the table below. Compare the two substances by accounting for the possible effect of the initial pressure on the decrease in pressure. Discuss the results of the ANCOVA analysis and compare them with those obtained by ignoring the potential effect of the initial pressure.

<table>
<thead>
<tr>
<th>Substance A</th>
<th>Substance B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal</td>
<td>Decrease</td>
</tr>
<tr>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>37</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
</tr>
</tbody>
</table>

We will use $\alpha = 0.05$ for all significance assessments.

```matlab
%x = Initial; y = Decrease; g=1 for 'A', g=2 for 'B'
x = [135 125 125 130 130 140 93 110 100 ... 90 135 130 115 110 140 95 90 105];
y = [45 45 20 50 25 37 50 20 25 15 ... 34 55 50 45 45 23 40 35];
g = [1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2];
a = 2; n = 10;
x1=x(g==1); x2=x(g==2);
y1=y(g==1); y2=y(g==2);
x1b=mean(x1); x2b=mean(x2);
y1b=mean(y1); y2b=mean(y2);
xb = mean(x); yb = mean(y);
SXX = sum( (x - xb).^2 );
SXY = sum( (x - xb).* (y - yb) );
SYY = sum( (y - yb).^2 );
QXX = sum( (x1-x1b).^2 + (x2-x2b).^2 );
```
\[ Q_{XY} = \text{sum}((x_1 - x_{1b}) \cdot (y_1 - y_{1b}) + \ldots + (x_2 - x_{2b}) \cdot (y_2 - y_{2b})); \]
\[ Q_{YY} = \text{sum}((y_1 - y_{1b})^2 + (y_2 - y_{2b})^2); \]
% estimators of model parameters \( \mu, \alpha_i, \beta \)
\[ \mu = yb \quad \%\mu = 36.7 \]
\[ b = \frac{Q_{XY}}{Q_{XX}} \quad \%b = 0.5175 \]
\[ \alpha = [y_{1b} \ y_{2b}] - [yb \ yb] - b \cdot ([x_{1b} \ x_{2b}] - [xb \ xb]) \]
\%\alpha = [-4.8714 \ 4.8714] \]
\[ T_{XX} = S_{XX} - Q_{XX}; \]
\[ T_{XY} = S_{XY} - Q_{XY}; \]
\[ T_{YY} = S_{YY} - Q_{YY}; \]
\[ \text{SSE} = Q_{YY} - \frac{Q_{XY}^2}{Q_{XX}}; \]
\%\text{MSE} = \frac{\text{SSE}}{a \cdot (n-1) - 1} \quad \%\text{MSE} = 60.6542 \]
\[ \text{SSEp} = S_{YY} - \frac{S_{XY}^2}{S_{XX}}; \]
% F-test for testing \( H_0: \alpha_i = 0 \) (all \( i \))
\[ F_1 = \frac{((\text{SSEp} - \text{SSE})/(a-1))}{/\text{MSE}} \]
\%\text{pvalF1} = 1 - fcdf(F_1, a-1, a \cdot (n-1) - 1) \quad %\text{pvalF1}=0.0133 \]
% F-test for testing \( H_0: \beta = 0 \)
\[ F_2 = \frac{(Q_{XY}^2/Q_{XX})}{/\text{MSE}} \]
\%\text{pvalF2} = 1 - fcdf(F_2, 1, a \cdot (n-1)-1) \]
\%\text{pvalF2} = 0.00012

Fig. 14.11 Scatterplot of decrease \( (y) \) against the initial value \( (x) \) for substances A and B.

Note that both null hypotheses are found to be significant, but the testing of the equality of treatments seems to be the more important hypothesis in this example. The conclusion is that substances A and B are significantly different at \( \alpha = 5\% \) significance level \( (p\text{-value} = 0.0133) \). Figure 14.11 plots...
the decrease \( y \) against the initial value \( x \) for substances A and B. The ANCOVA fits \( \hat{y} = 36.7 \pm 4.8714 + 0.5175 \cdot (x - 116.65) \) are superimposed.

If we conducted the ANOVA test by ignoring the covariates, the test would find no differences among the drugs (\( p \)-value = 0.2025). Likewise, if both treatments were lumped together and the response regressed on \( x \), the regression would be significant but with a \( p \)-value more than eight times larger than in ANCOVA.

\[
oX = [\text{ones}(\text{size}(x')) \ x'];\n[b, bint, r, rint, stats] = \text{regress}(y', oX);\n\text{stats} \% \; 0.4599 \; 15.3268 \; 0.0010 \; 83.0023\n[p, table, statan] = \text{anova1}(y, g);\np \% 0.2025
\]

This example shows how important accounting for sensible predictors is and how selecting the right statistical procedure is critical in modeling and decision making.

MATLAB has a built-in function \texttt{aoctool} that, when invoked as \texttt{aoctool(x,y,g)}, opens a front-end suite with various modeling and graphing capabilities. The output \texttt{stats} from \texttt{aoctool} can be imported into \texttt{multcompare} for a subsequent pairwise comparison analysis.

### 14.11.1 Sample Size in ANCOVA

ANCOVA can be thought of as an ANOVA for the transformed observations,

\[
y_{ij}^* = y_{ij} - \beta(x_{ij} - \bar{x}), \quad i = 1, 2, \ldots, a; \quad j = 1, \ldots, n.
\]

where, as before, \( a \) denotes the number of groups/treatments, and \( n \) is the common group size. The total sample size is \( N = an \).

Thus the power analysis in ANCOVA is related to that of ANOVA with a small modification. Namely, for a single covariate the number of degrees of freedom in the denominator of \( F \) statistic becomes \( (n - 1)a - 1 \) instead of \( (n - 1)a \). More generally, if \( k \) covariates are employed, the number of degrees of freedom would reduce to \( (n - 1)a - k \).

With this modification power analysis proceeds as it was described in Section 11.8.

**Example 14.9. Kodlin’s Experiment Revisited.** If we are to repeat Kodlin’s experiment described in Example 14.8 and wanted to be able to find medium effect \( f^2 = 0.25^2 \) with the power of 85% in \( \alpha = 0.05 \) level testing, we will use (11.5). Here we have a single covariate, \( a = 2 \) treatments/groups, and \( n = 10 \) observations per group. The power for such experiment would be 0.1872.
Thus, the experimenters “were lucky” that the observed effect in the experiment was large enough to produce a significant result (p-value was 0.0133).

If we were to design a new experiment and require a power of at least 85%, for the medium effect \( f^2 = 0.25^2 \), and \( \alpha = 0.05 \), the required sample size would be 73 per treatment.

\[
a=2 ; \quad f^2=0.25^2 ; \\
f= @(n) 1-ncfcdf( finv(0.95, a-1, (n-1)*a - 1), a-1, (n-1)*a - 1, a*n*f2)-0.85 \\
ssize=fzero(f,100) \quad %ssize =72.8073
\]

### 14.11.2 Bayesian Approach to ANCOVA

Bayesian approach implemented via WinBUGS is conceptually simple: the ANCOVA model is stated in its direct form, constraints on the ANOVA coefficients imposed (either STZ or CR), and the priors on all free parameters are elicited and set. Unlike the classical approach where more than one covariate in the regression part and more than one factor in the ANOVA part significantly increase the complexity of ANCOVA model, the Bayesian approach via WinBUGS remains simple and straightforward.

The following example solved in WinBUGS illustrates handling ANCOVA in a Bayesian manner.

**Example 14.10. ANCOVA Fibers.** Three machines produce monofilament synthetic fiber for medical use (surgery, implants, devices, etc.). The measured response is the strength \( y \) (in pounds, lb.) and a covariate is the diameter \( x \) (in inches/1000).

```r
#ancovafibers.odc

model{
  for (i in 1:ntotal){
    y[i] ~ dnorm( mui[i], tau )
    mui[i] <- mu + alpha[g[i]] + beta1 *(x[i] - mean(x[]))
  }

  #alpha[1] <- 0.0; #CR constraints

  mu ~ dnorm(0, 0.0001)
  beta0 ~ dnorm(0, 0.0001)
  beta0 <- mu - beta1 * mean(x[])
```
14.11 Analysis of Covariance

\[ \text{alpha}[2] \sim \text{dnorm}(0, 0.0001) \]
\[ \text{alpha}[3] \sim \text{dnorm}(0, 0.0001) \]
\[ \tau \sim \text{dgamma}(0.001, 0.001) \]
\[ \text{var} \leftarrow 1/\tau \]

\textbf{DATA}
\begin{verbatim}
list(ntotal=15, a=3,
y = c(36, 41, 39, 42, 49,
    40, 49, 39, 45, 44,
    35, 37, 42, 34, 32),
x = c(20, 25, 24, 25, 32,
    22, 28, 22, 30, 28,
    21, 23, 26, 21, 15),
g = c(1, 1, 1, 1, 1,
    2, 2, 2, 2, 2,
    3, 3, 3, 3, 3) )
\end{verbatim}

\textbf{INITS}
\begin{verbatim}
list(mu=1, alpha=c(NA, 0, 0), beta1=0, tau=1)
\end{verbatim}

The output from WinBUGS is

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>val2.5pc</th>
<th>median</th>
<th>val97.5pc</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha[1]</td>
<td>0.182</td>
<td>0.6618</td>
<td>0.001848</td>
<td>-1.129</td>
<td>0.1833</td>
<td>1.494</td>
<td>10001</td>
<td>100000</td>
</tr>
<tr>
<td>alpha[2]</td>
<td>1.218</td>
<td>0.6888</td>
<td>0.003279</td>
<td>-0.1576</td>
<td>1.217</td>
<td>2.596</td>
<td>10001</td>
<td>100000</td>
</tr>
<tr>
<td>alpha[3]</td>
<td>-1.4</td>
<td>0.743</td>
<td>0.003757</td>
<td>-2.876</td>
<td>-1.399</td>
<td>0.07178</td>
<td>10001</td>
<td>100000</td>
</tr>
<tr>
<td>beta0</td>
<td>17.18</td>
<td>3.086</td>
<td>0.01394</td>
<td>11.01</td>
<td>17.18</td>
<td>23.31</td>
<td>10001</td>
<td>100000</td>
</tr>
<tr>
<td>beta1</td>
<td>0.9547</td>
<td>0.1266</td>
<td>5.634E-4</td>
<td>0.7019</td>
<td>0.9543</td>
<td>1.207</td>
<td>10001</td>
<td>100000</td>
</tr>
<tr>
<td>mu</td>
<td>40.2</td>
<td>0.4564</td>
<td>0.001372</td>
<td>39.29</td>
<td>40.2</td>
<td>41.11</td>
<td>10001</td>
<td>100000</td>
</tr>
<tr>
<td>var</td>
<td>3.118</td>
<td>1.671</td>
<td>0.007652</td>
<td>1.276</td>
<td>2.716</td>
<td>7.348</td>
<td>10001</td>
<td>100000</td>
</tr>
</tbody>
</table>

The regression equations corresponding to the three treatments (machines) are

\[ \hat{y}_{1i} = 17.18 + 0.182 + 0.9547 \cdot x_i, \]
\[ \hat{y}_{2i} = 17.18 + 1.218 + 0.9547 \cdot x_i, \text{ and} \]
\[ \hat{y}_{3i} = 17.18 - 1.4 + 0.9547 \cdot x_i. \]

Note that all three 95% credible sets for alpha contain 0, while the credible set for the slope beta1 does not contain 0. This analysis of credible sets is not a formal Bayesian testing, but it agrees with the output from

\[ y = [36, 41, 39, 42, 49, 40, 48, 39, 45, 44, 35, 37, 42, 34, 32]; \]
\[ x = [20, 25, 24, 25, 32, 22, 28, 22, 30, 28, 21, 23, 26, 21, 15]; \]
\[ g = [1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3]; \]
\[ \text{aoctool}(x, y, g, \text{parallel lines option}) \]

where the p-value corresponding to ANOVA part is 0.11808. The test for the regression slope is significant with a p-value of 4.2645e-06.
14.12 Exercises

14.1. **Regression with Three Points.** Points \((x_1, y_1) = (1, 1), (x_2, y_2) = (2, 2),\) and \((x_3, y_3) = (3, 2)\) are given.
(a) Find (by hand calculation) the regression line that best fits the points, and sketch a scatterplot of points with the superimposed fit.
(b) What are the predictions at 1, 2, and 3 (i.e, \(\hat{y}_1, \hat{y}_2,\) and \(\hat{y}_3)\)?
(c) What are \(\text{SST}, \text{SSR},\) and \(\text{SSE}\)?
(d) Find an estimator of variance \(\hat{\sigma}^2\).

14.2. **Age and IVF Success Rate.** The highly publicized (recent TV reports) in vitro fertilization (IVF) success cases for women in their late fifties all involve donors’ eggs. If the egg is the woman’s own, the story is quite different.
IVF, an assisted reproductive technology (ART) procedure, involves extracting a woman’s eggs, fertilizing the eggs in the laboratory, and then transferring the resulting embryos to the woman’s uterus through the cervix. Fertilization involves a specialized technique known as intracytoplasmic sperm injection (ICSI).
The table below shows the live-birth success rate per transfer rate from a woman’s own eggs, by age of recipient. The data are for the year 1999, published by the CDC at [http://www.cdc.gov/art/ARTReports.htm](http://www.cdc.gov/art/ARTReports.htm).

<table>
<thead>
<tr>
<th>Age (x)</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage (y)</td>
<td>38.7</td>
<td>38.6</td>
<td>38.9</td>
<td>41.4</td>
<td>39.7</td>
<td>41.1</td>
<td>38.7</td>
<td>37.6</td>
<td>36.3</td>
<td>36.9</td>
<td>35.7</td>
<td>33.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age (x)</th>
<th>36</th>
<th>37</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>41</th>
<th>42</th>
<th>43</th>
<th>44</th>
<th>45</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage (y)</td>
<td>33.2</td>
<td>30.1</td>
<td>27.8</td>
<td>22.7</td>
<td>21.3</td>
<td>15.4</td>
<td>11.2</td>
<td>9.2</td>
<td>5.4</td>
<td>3.0</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Select the ages in the range 33–46, which shows an almost linear decay of the success rate (Figure 14.12). For the selected ages fit the linear model \(\hat{y} = b_0 + b_1 x\). Would the quadratic relationship \(\hat{y} = b_0 + b_1 x + b_2 x^2\) be more appropriate than the linear?

14.3. **Sharp Dissection and Severity of Postoperative Adhesions.** Postoperative adhesions are formed after surgical cardiac and great vessel procedures as part of the healing process. Scar tissue makes reentry complex and increases the rate of iatrogenic lesions. Currently, as reoperations are needed in 10% to 20% of heart surgeries, various methods have been investigated to prevent or decrease the severity of postoperative adhesions.
The surgical time spent in the adhesiolysis procedure and amounts of sharp dissection are informative summaries to to predict the severity of pericardial adhesions, as reported by Lopes et al. (2009). For example, the authors reported a linear relationship between the logarithm of the amount of sharp dissection \(\text{lasd}\) and severity score \(\text{sesco}\) assessed by a
standard categorization. Use the data in MATLAB format to answer (a)–(e).

\[ shdiss = [14 \ 108 \ 39 \ 311 \ 24 \ 112 \ 104 \ldots \\
382 \ 42 \ 74 \ 67 \ 145 \ 21 \ 93 \ldots \\
75 \ 381 \ 36 \ 36 \ 73 \ 239 \ 35 \ 69]; \]

\[ \text{lasd} = \log(\text{shdiss}); \]

\[ \text{sesco} = [6 \ 12 \ 9 \ 18 \ 7 \ 12 \ 12 \ldots \\
17 \ 8 \ 11 \ 14 \ 15 \ 7 \ 12 \ldots \\
13 \ 18 \ 9 \ 9 \ 10 \ 16 \ 7 \ 10]; \]

(a) Write down the linear relationship between \text{lasd} and \text{sesco}.
(b) What is $R^2$ here and what does it represent?
(c) Test the hypothesis $H_0 : \beta_0 = 0$ versus the alternative $H_1 : \beta_0 < 0$. Use $\alpha = 0.05$. The critical cut points are provided at the back of the problem, or, alternatively, you can report the $p$-value.
(d) Find a 95% confidence interval for the population slope $\beta_1$.
(e) For \text{lasd=4} predict the severity score. Find a 99% confidence interval for the mean response.

14.4. **Mg–CaO Data Revisited.** Consider the data on CaO measurements discussed in Example 13.2. In data matrix hazel.dat|mat there are columns $y$ and $z$ corresponding to two ways to assess CaO (methods A and B), as well as a column $x$ that gives the exact CaO amount.
(a) Fit two linear regressions. The first regression should express CaO found (A) ($y$) as a linear function of CaO present ($x$); that is, find the

---

**Fig. 14.12** Success rate (in %) versus age (in years).
equation \( y = b_0 + b_1 x \). Then find the second regression as \( z = d_0 + d_1 x \). Discuss the adequacy of regression fits.

(b) Test the hypothesis about the slope for the regression for (A) \( H_0 : \beta_1 = 1 \) versus the alternative \( H_1 : \beta_1 < 1 \).

(c) Find a 95% confidence interval (CI) for the population intercepts for each of the regressions. Is 0 in any of the intervals? What does it mean if it is and if it is not?

(d) Find the average CaO found by method A if the CaO present was \( x^\ast = 15 \).

(e) A small sample is sent to your lab to be analyzed by method A. It is known that the CaO present in the sample is \( x^\ast = 15 \). What is the 95% prediction interval (PI) for the CaO found?

14.5. Kanamycin Levels in Premature Babies. Miller (1980) describes a project and provides data on assessing the precision of noninvasive measuring of kanamycin concentration in neonates. Premature babies are susceptible to infections, and kanamycin (an aminoglycoside) is used for the treatment of sepsis. Since kanamycin is ineffective at low doses and potentially harmful at high doses, it is necessary to constantly monitor its levels in a premature baby’s body during treatment. The standard procedure for measuring serum kanamycin levels is to take blood samples from a heel. Unfortunately, due to frequent blood sampling, neonates are left with badly bruised heels. Kanamycin is routinely administered through an umbilical catheter. An alternative procedure for measuring serum kanamycin would be to reverse the flow in the catheter and draw a blood sample from it. The concern about this noninvasive method is that the blood drawn from the point close to an infusion may have an elevated level of kanamycin compared to blood samples from more distant points in the body. In a carefully designed experimental setup, blood samples from 20 babies were obtained simultaneously from an umbilical catheter and a heel venapuncture (using a heelstick). If the agreement is satisfactory, physicians would be willing to use the catheter values instead of heelstick values.

Here are the data:
(a) Model the Heelstick responses with Catheter as the predictor in a linear regression.

(b) Are there any unusual observations? Does regression improve when unusual observations are removed from the analysis?

(c) Test $H_0 : \beta_1 = 1$ versus $H_1 : \beta_1 < 1$ at the level $\alpha = 0.05$.

(d) Find a 97% confidence interval for the population intercept.

(e) Find 95% confidence and prediction intervals for the regression response when $cath = 20$.

(f) Using WinBUGS, estimate the parameters in a Bayesian regression with noninformative priors. Compare Bayesian and the least-squares solutions.

14.6. Degradation of Scaffolds. In an experiment conducted at the Georgia Tech/Emory Center for the Engineering of Living Tissues, the goal was to find a suitable biomechanical replacement for cartilage, better known as tissue engineered cartilage. There are many factors (dimensional or mechanical) at which the cartilage scaffold is tested to assess whether it is a viable replacement. One of the problems is the degradation of scaffolds as the tissue grows, which affects all of the experimental metrics. The experimental data collected comprise a tissue growth experiment in which no cells were added, thus approximating the degradation of the scaffold over a sequence of 8 days. The dynamic shear summaries capture two physical phenomena, the modulus or the construct’s ability to resist deformation under load and the frequency at which the modulus was evaluated. This modulus provides a measure of the extent of interconnectivity within the fibrous scaffold.

The table below contains moduli, in 1000s, for the frequency $f = 1$ over 8 days, each day represented by three independent measurements.
(a) Use the moduli for frequency $f = 1$. Write down the linear regression model: $mod1 = b_0 + b_1 \cdot \text{day}$, where $b_0$ and $b_1$ are estimators of the population intercept and slope (Figure 14.13). What is $R^2$ for your regression?

(b) Test the hypothesis that the population intercept is equal to 100 versus the alternative that it is smaller than 100.

(c) Find a 96% confidence interval for the population slope.

(d) For $\text{day} = 5.5$, find the prediction of the modulus. What is the standard deviation for this predicted value?

14.7. **Glucosis in Lactococcus Lactis.** The data set Lactis.dat is courtesy of Dr. Eberhard Voit at the Georgia Institute of Technology and is an excerpt from a larger collection of data dealing with glycolysis in the bacterium *Lactococcus lactis* MG1363 (which is involved in essentially all yogurts, cheeses, etc.). The experiment was conducted in aerobic conditions with cell suspension in a 50-mM KPi buffer with a pH of 6.5, and 20 mM...
14.12 Exercises

[6-13C] glucose. There are four columns in the file `Lactis.dat`: time, glucose, lactate, and acetate.
The levels of extracellular metabolites lactate and acetate are monitored over time. After time \( t = 5 \), the level of lactate stabilizes around 30, while the level of acetate shows a linearly increasing trend in time.
(a) Plot lactate level against time. Take a subset of lactate levels for times \( t > 5 \) (86 observations) and find basic descriptive statistics.
(b) Check the normality of the subset data. Test the hypothesis that the mean lactate level for \( t > 5 \) is 30 against the two-sided alternative.
(c) Select acetate levels for \( t > 5 \). Fit a linear relationship of acetate level against time and show that the linear model is justified by providing and discussing the ANOVA table.

14.8. **Weight and Latency in Rats.** Data consisting of rat body weight (grams) and latency to seizure (minutes) are given for 15 rats (adapted from Kleinbaum et al., 1987):

<table>
<thead>
<tr>
<th>Rat #</th>
<th>Weight</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>348</td>
<td>1.80</td>
</tr>
<tr>
<td>2</td>
<td>372</td>
<td>1.95</td>
</tr>
<tr>
<td>3</td>
<td>378</td>
<td>2.90</td>
</tr>
<tr>
<td>4</td>
<td>390</td>
<td>2.30</td>
</tr>
<tr>
<td>5</td>
<td>392</td>
<td>1.10</td>
</tr>
<tr>
<td>6</td>
<td>395</td>
<td>2.50</td>
</tr>
<tr>
<td>7</td>
<td>400</td>
<td>1.30</td>
</tr>
<tr>
<td>8</td>
<td>409</td>
<td>2.00</td>
</tr>
<tr>
<td>9</td>
<td>413</td>
<td>1.70</td>
</tr>
<tr>
<td>10</td>
<td>415</td>
<td>2.95</td>
</tr>
<tr>
<td>11</td>
<td>423</td>
<td>2.25</td>
</tr>
<tr>
<td>12</td>
<td>428</td>
<td>2.25</td>
</tr>
<tr>
<td>13</td>
<td>464</td>
<td>3.05</td>
</tr>
<tr>
<td>14</td>
<td>468</td>
<td>3.70</td>
</tr>
<tr>
<td>15</td>
<td>470</td>
<td>3.62</td>
</tr>
</tbody>
</table>

It is of interest to regress the latency \( y \) to weight \( x \).
(a) Test the hypothesis \( H_0 : \beta_0 = 0 \) against the two-sided alternative. Use \( \alpha = 0.05 \).
(b) Test the hypothesis \( H_0 : \beta_1 = 0.02 \) against the alternative \( H_0 : \beta_1 < 0.02 \). Use \( \alpha = 0.05 \).
(c) Find a 95% confidence interval for the slope \( \beta_1 \).
(d) For weight \( x = 410 \) find the mean latency response, \( \hat{y}_m \). Test the hypothesis \( H_0 : \hat{y}_m = 3 \) versus the alternative \( H_1 : \hat{y}_m < 3 \). Test the same hypothesis for the predicted response \( \hat{y}_{pred} \). In both tests use \( \alpha = 0.05 \).

14.9. **Rinderpest Virus in Rabbits.** Temperatures \( \text{temp} \) were recorded in a rabbit at various times \( \text{time} \) after the rabbit was inoculated with rinderpest virus (modified from Carter and Mitchell, 1958). Rinderpest (RP) is an infectious viral disease of cattle, domestic buffalo, and some species of wildlife; it is commonly referred to as cattle plague. It is characterized by fever, oral erosions, diarrhea, lymphoid necrosis, and high mortality.
(a) Demonstrate that a linear regression with one predictor (time) gives an insignificant $F$-statistic and relatively low $R^2$.
(b) Include time$^2$ (squared time) as the second predictor, making the regression quadratic in variables, but still linear in coefficients. Show that this regression is significant and has a larger $R^2$.
(c) Find the 90% confidence interval for the coefficient of the quadratic term and test the hypothesis that the intercept is equal to 100 versus the one-sided alternative that it is smaller than 100. Use $\alpha = 0.05$.

14.10. **Hemodilution.** Clark et al. (1975) examined the fat filtration characteristics of a packed polyester-and-wool filter used in arterial lines during clinical hemodilution. They collected data on the filter’s recovery of solids for ten patients who underwent surgery. The table below shows removal rates of lipids and cholesterol:

<table>
<thead>
<tr>
<th>Patient</th>
<th>Lipids ($x$)</th>
<th>Cholesterol ($y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.81</td>
<td>1.90</td>
</tr>
<tr>
<td>2</td>
<td>2.10</td>
<td>1.03</td>
</tr>
<tr>
<td>3</td>
<td>0.79</td>
<td>0.44</td>
</tr>
<tr>
<td>4</td>
<td>1.99</td>
<td>1.18</td>
</tr>
<tr>
<td>5</td>
<td>1.03</td>
<td>0.62</td>
</tr>
<tr>
<td>6</td>
<td>2.07</td>
<td>1.29</td>
</tr>
<tr>
<td>7</td>
<td>0.74</td>
<td>0.39</td>
</tr>
<tr>
<td>8</td>
<td>3.88</td>
<td>2.30</td>
</tr>
<tr>
<td>9</td>
<td>1.43</td>
<td>0.93</td>
</tr>
<tr>
<td>10</td>
<td>0.41</td>
<td>0.29</td>
</tr>
</tbody>
</table>

(a) Fit a regression line to the data, with cholesterol as the response variable and lipids as the covariate. Discuss the adequacy of the proposed linear fit.
(b) What test is the resulting regression $p$-value referring to? State $H_0$ and $H_1$.
(c) Find the 95% confidence interval for the population intercept $\beta_0$.
(d) Test the hypothesis that the population slope $\beta_1$ is equal to 2/3 versus the one-sided alternative that it is less than 2/3. Use $\alpha = 0.05$. 

<table>
<thead>
<tr>
<th>Time after injection</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>(time in hrs)</td>
<td>(temp in °F)</td>
</tr>
<tr>
<td>24</td>
<td>102.8</td>
</tr>
<tr>
<td>32</td>
<td>104.5</td>
</tr>
<tr>
<td>48</td>
<td>106.5</td>
</tr>
<tr>
<td>56</td>
<td>107.0</td>
</tr>
<tr>
<td>70</td>
<td>105.1</td>
</tr>
<tr>
<td>72</td>
<td>103.9</td>
</tr>
<tr>
<td>80</td>
<td>103.2</td>
</tr>
<tr>
<td>96</td>
<td>102.1</td>
</tr>
</tbody>
</table>
14.12 Exercises 741

(e) Predict the cholesterol rate for lipids at the level $1.65 \text{ mg/kg/L} \times 10^{-2}$. Find the 95% confidence interval for the mean response and prediction interval for individual response.

14.11. Anscombe’s Data Sets. A celebrated classic example of the role of residual analysis and statistical graphics in statistical modeling was created by Anscombe (1973). He constructed four different data sets $(x_i, y_i), i = 1, \ldots, 11$, that share the descriptive statistics necessary to establish a linear regression fit $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

A linear model is appropriate for data set 1; the scatterplots and residual analysis suggest that data sets 2–4 seem not to be amenable to linear modeling.

(a) Using MATLAB fit the regression line for the four sets and provide four ANOVA tables. What statistics are the same? What statistics differ?
(b) Plot the residuals against the fitted values. Discuss the appropriateness of the regressions.

14.12. Potato Leafhopper. Potato leafhopper (Empoasca fabae) is found throughout much of the United States east of the Rocky Mountains. It feeds on nearly 200 kinds of plants. Feeding and egg laying cause the infested plant damage. Eggs are deposited in the midrib or larger veins of the leaves, or in the petioles or stems. The leaves turn yellow, or sometimes pink or purple, and become wilted or stunted.

The length of the developmental period (in days) of the potato leafhopper, from egg to adult seems to be dependent on temperature (Kouskolekas and Decker, 1966). The original data were weighted means, but for the purpose of this exercise we consider them as though they were single observed values.

(a) Find a 98% confidence interval for population slope $\beta_1$. 
(b) Test the hypothesis that the intercept is equal to 60 against the alternative that it is larger than 60. Take $\alpha = 0.01$.

(c) What is the 96% confidence interval for the mean response (mean number of days) if the temperature is 85°F?

14.13. **Force Sensor Calibration.** Your friend is conducting an experiment to measure the grasping force a robot arm exerts on an object. To accomplish this, he puts a force sensor on an object so that when the robot arm grasps it, he can measure the force. The force sensor maps change in electric pressure (in Volts) to force (in Newtons).

![Image](Fig. 14.14 The robot arm grasps on an object that has a force sensor. The force sensor outputs voltage readings which can be mapped to force.)

The more you push on the force sensor, the higher is the electric pressure, and the higher is the voltage reading. A setup is shown in Figure 14.14. Before starting the experiment, your friend runs into a problem: he lost the force sensor data sheet! To calibrate the force sensor, he applies known weights to the force sensor and records the voltage output. His measurements are shown below:

<table>
<thead>
<tr>
<th>Force ($x$)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voltage ($y$)</td>
<td>326</td>
<td>375</td>
<td>403</td>
<td>438</td>
<td>555</td>
<td>646</td>
<td>799</td>
<td>1005</td>
<td>1223</td>
<td>1383</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Force ($x$)</th>
<th>30</th>
<th>60</th>
<th>65</th>
<th>70</th>
<th>80</th>
<th>100</th>
<th>110</th>
<th>130</th>
<th>150</th>
<th>170</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voltage ($y$)</td>
<td>1653</td>
<td>2207</td>
<td>2250</td>
<td>2400</td>
<td>2767</td>
<td>3287</td>
<td>3727</td>
<td>4513</td>
<td>5170</td>
<td>5417</td>
</tr>
</tbody>
</table>

(a) Given the data he collected, fit a linear regression with force as the independent, controlled variable and voltage as the response.

(b) When the robot grasps the object, your friend records a voltage of 850 Volts. Using the regression from (a) estimate what this voltage corresponds to in force? Denote this force by $x^*$ and its estimator by $\hat{x}^*$. Find a 95% CI for $x^*$.

(c) Flip now the roles of $x$ and $y$ and fit a linear regression in which $x$ (force) is predicted by $y$ (voltage). Predict now the force $x^*$ for a voltage of 850 and compare it with the prediction in (b). Compare 95% CI’s as well.
(d) Although the prediction in part (c) looks “more natural,” it is not formally correct; the prediction as in (b) is recommended. Can you guess why? State your arguments in one or two sentences.

14.14. **Determining Average Slope.** One way to learn about the function of the kidney is to study the rate at which it produces and consumes different substances. An important quantity is the rate at which oxygen is consumed, since this is related to the kidney’s workload. Another item of interest is the rate at which the kidney reabsorbs ionic sodium from the urine. This activity, known as sodium pumping, requires energy.

Data analyzed by Hyde (1980) contains measurements from 8 kidneys treated by the drug ouabain. Ouabain, from Somali waabayo – “arrow poison,” is g-strophanthin, a poisonous cardiac glycoside. This drug inhibits sodium pumping in kidneys.

The first column is the number of the kidney, the second column provides dry weight in milligrams multiplied by 10, the third column shows sodium reabsorption in $\mu$Eq/min, multiplied by 100. The last column is oxygen consumption in $\mu$moles/min, multiplied by 1000. The measurements are taken over 10-minute intervals.

Although both measurements have errors, the investigators felt that sodium measurements are more precise and should be taken as the covariate, and oxygen measurements as the response.

The average slope in regressions is of interest; in fact researchers were interested in estimating the reciprocal of the slope, which measures sodium pumping efficiency of the kidney.

(a) Find the slopes $b_{1,i}, i = 1, \ldots, 8$, in regressions of oxygen consumption (as $y$) to sodium reabsorption (as $x$), one for each kidney.

(b) Find a weighted average $b_1 = \frac{\sum_{i=1}^{8} w_i b_{1,i}}{\sum_{i=1}^{8} w_i}$, where the weights $w_i$ are proportional to $1/\sum_{i=1}^{8} w_i$ and sum up to 1. Here $S_{xx,i}$ is $\sum_{k=1}^{6} (x_k - \bar{x})^2$ for the $i$th kidney. What is the variance of $b_1$?

(c) Using Bayesian hierarchical model, estimate an overall slope. Use noninformative priors. Compare this slope and its variance to results from (b).

**Hint:** In (b) the weights are inverse proportional to slope variances. The variance of weighted average slope is $\sigma^2 / \sum_{i=1}^{8} S_{xx,i}$. For part (c) consult Rats from Examples Vol I provided with the distribution of WinBUGS/OpenBUGS.

14.15. **Cross-validating a Bayesian Regression.** In this exercise covariates $x_1$ and $x_2$ are simulated as

\[
x_1 = \text{rand}(1, 40) \quad \text{and} \quad x_2 = \text{floor}(10 \times \text{rand}(1,40)) + 1;
\]

and the response variable $y$ is obtained as

\[
y = 2 + 6 \times x_1 - 0.5 \times x_2 + 0.8 \times \text{randn(size(x1))};
\]
Write a WinBUGS program that selects 20 triples \((x_1, x_2, y)\) to train the linear regression model \(\hat{y} = b_0 + b_1 x_1 + b_2 x_2\) and then uses the remaining 20 triples to evaluate the model by comparing the original responses \(y_i, i = 21, \ldots, 40\), with regression-predicted values \(\hat{y}_i, i = 21, \ldots, 40\). The comparison involves calculating the MSE, the mean of \((y_i - \hat{y}_i)^2, i = 21, \ldots, 40\).

This is an example of how a cross-validation methodology is often employed to assess statistical models.

How do the Bayesian estimators of \(\beta_0, \beta_1, \beta_2,\) and \(\sigma\) compare to the “true” values 2, 6, −0.5, and 0.8?

14.16. **Taste of Cheese.** As cheddar cheese matures, a variety of chemical processes take place. The taste of mature cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and were subjected to taste tests. The table below presents data (from the experiments of G. T. Lloyd and E. H. Ramshaw, CISRO Food Research, Victoria, Australia, analyzed in Moore and McCabe, 2006) for one type of cheese manufacturing process. **Taste** is the response variable of interest. The taste scores were obtained by combining scores from several tasters. Three of the chemicals whose concentrations were measured are acetic acid, hydrogen sulfide, and lactic acid. For acetic acid and hydrogen sulfide, log transformations were taken.

<table>
<thead>
<tr>
<th>Taste</th>
<th>Acetic</th>
<th>H2S</th>
<th>Lactic</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.3</td>
<td>4.54</td>
<td>3.13</td>
<td>0.86</td>
</tr>
<tr>
<td>39.0</td>
<td>5.37</td>
<td>5.44</td>
<td>1.57</td>
</tr>
<tr>
<td>5.6</td>
<td>4.66</td>
<td>3.81</td>
<td>0.99</td>
</tr>
<tr>
<td>37.3</td>
<td>5.89</td>
<td>8.73</td>
<td>1.29</td>
</tr>
<tr>
<td>18.1</td>
<td>4.90</td>
<td>3.85</td>
<td>1.29</td>
</tr>
<tr>
<td>34.9</td>
<td>5.74</td>
<td>6.14</td>
<td>1.68</td>
</tr>
<tr>
<td>0.7</td>
<td>4.48</td>
<td>3.00</td>
<td>1.06</td>
</tr>
<tr>
<td>54.9</td>
<td>6.15</td>
<td>6.75</td>
<td>1.52</td>
</tr>
<tr>
<td>15.9</td>
<td>4.79</td>
<td>3.91</td>
<td>1.16</td>
</tr>
<tr>
<td>18.0</td>
<td>5.25</td>
<td>6.17</td>
<td>1.63</td>
</tr>
<tr>
<td>14.0</td>
<td>4.56</td>
<td>4.95</td>
<td>1.15</td>
</tr>
<tr>
<td>32.0</td>
<td>5.46</td>
<td>9.24</td>
<td>1.44</td>
</tr>
<tr>
<td>16.8</td>
<td>5.37</td>
<td>3.66</td>
<td>1.31</td>
</tr>
<tr>
<td>26.5</td>
<td>6.46</td>
<td>6.92</td>
<td>1.72</td>
</tr>
<tr>
<td>13.4</td>
<td>5.80</td>
<td>6.69</td>
<td>1.08</td>
</tr>
</tbody>
</table>

(a) Find the equation in multiple linear regression that predicts **Taste** using **Acetic**, **H2S**, and **Lactic** as covariates.

(b) For **Acetic** = 5, **H2S** = 8, and **Lactic** = 2, estimate the regression response \(\hat{Y}_b\) and find standard deviations for the mean and individual responses. [Ans. 43.37, 6.243, 11.886]
14.12 Exercises

(c) Find the 98% confidence interval for the intercept $\beta_0$.
(d) Construct an ANOVA table.
(e) Find ordinary, studentized, and studentized deleted residuals for the observation $Y_8$.
(f) Find $DFITS_8$ and $COOKSD_8$.
(g) Find $DFBETAS_8$ on the Lactic coefficient.
(h) Find and discuss the VIF.

14.17. Slowing the Progression of Arthritis. Arthritis is caused by the breakdown of collagen in joint cartilage by the enzyme MMP-13. The antibiotic doxycycline is a general inhibitor of MMPs and, by inhibiting the activity of MMP-13, is an effective method of slowing the progression of arthritis. At present, doxycycline is used to treat both rheumatoid arthritis and osteoarthritis. The rabbit’s HIG-82 synovial cell line was used to model arthritis. MMP-13 was prepared by adding PMA, which guarantees its presence, and APMA, which activates it. The enzyme MMP-13 was mixed with a quenched substrate. When the enzyme cleaves the substrate, it fluoresces. This fluorescence was used to measure the amount of MMP-13 activity. Doxycycline, which decreases the amount of enzyme activity, was added in increasing concentrations: 0, 25, 50, 75, 100, and 200 micromols. The decrease in the fluorescence produced by the cleaved substrate when doxycycline was present was used to measure the decrease in activity of the enzyme. The same experiment was performed on three different plates, and the data was normalized.

The data set arthritis1.mat can be found on the book’s website. The data file contains 72 observations (rows); the first column represents doxycycline concentration (0, 25, 50, 75, 100, and 200), the second column is the fluorescence response, and the third column is the plate number.

(a) Fit the linear regression where fluorescence is the response and doxycycline concentration is the predictor. Predict the fluorescence if the doxycycline concentration is equal to 125.
(b) Since three plates are present, run aoctool with doxycycline concentration as $x$, fluorescence as $y$, and plate number as $g$. Is there a significant difference between the plates?

14.18. Drosophila Offspring Prediction. In a Genetics Lab at Duke University students were involved in a project aimed to predict the sizes of Drosophila offspring population.

For 22 different mating bottles, a number of yeast flakes and a number of virgin flies was set and recorded. After several days the number of offsprings was counted.

For all bottles, the genotype of flies, temperature, and the number of males (3 per bottle), were kept the same. The bottles were attended in the same manner and the count was made at 4:00 pm for each bottle. The

---

4 David Lee and Craig Cook, Project in STAT110, ISDS, Duke University, Spring 1996.
earliest counting day was after 2 weeks of setting the bottle, since virgin flies need at least 13-14 days to reproduce after mating. Data is available in the file drosophila.dat|mat|xlsx.

<table>
<thead>
<tr>
<th>Bottle</th>
<th>Flakes</th>
<th>Virgins</th>
<th>Days</th>
<th>Offsprings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>11</td>
<td>23</td>
<td>75</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>9</td>
<td>20</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>7</td>
<td>17</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>15</td>
<td>18</td>
<td>67</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>9</td>
<td>17</td>
<td>39</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>6</td>
<td>25</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>10</td>
<td>20</td>
<td>45</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>7</td>
<td>16</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>11</td>
<td>23</td>
<td>75</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>10</td>
<td>18</td>
<td>61</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>1</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>5</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>4</td>
<td>21</td>
<td>35</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>12</td>
<td>22</td>
<td>72</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>12</td>
<td>24</td>
<td>85</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>11</td>
<td>22</td>
<td>51</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
<td>6</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>8</td>
<td>21</td>
<td>50</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>9</td>
<td>21</td>
<td>61</td>
</tr>
<tr>
<td>21</td>
<td>5</td>
<td>7</td>
<td>24</td>
<td>48</td>
</tr>
<tr>
<td>22</td>
<td>8</td>
<td>10</td>
<td>19</td>
<td>51</td>
</tr>
</tbody>
</table>

Although counts of offsprings are integers and Poisson regression is appropriate (to be covered in Chapter 15), since the average offspring count is about 50, normal approximation is very good and the multiple linear regression is adequate.

(a) Regress offsprings on flakes, virgins, and days. Write down the equation that links the number of offsprings to predictors.

(b) What is the overall $R^2$? A new bottle is set with 10 flakes and 10 virgins. How many offsprings is predicted if the count is made after 22 days?

(c) Run MATLAB’s procedure stepwise. Is a smaller model recommended?

(d) Variable virgins enters during stepwise procedure first and accounts for more than 66% of total variability in the response. How many offsprings do you predict using univariate regression with only this variable as the predictor (the same value as in (b): virgins=10)? Find a 95% Prediction Interval for the number of offsprings in this case. Is the offspring prediction from (b) contained in this interval?

14.19. **Insulin on Opossum Liver.** Corkill (1932) provides data on the influence of insulin on opossum liver. In the experimental setup the 20 animals (common gray Australian opossums – *Trichosurus*) fasted for 24 or 36 hours. Ten animals, four from the 24-hour fasting group and six from the 36-hour fasting group, were injected with insulin, while the remaining ten animals served as controls, that is, they received no insulin. After 3 to 4 hours liver glycogen and blood sugar were measured. The weights of the animals were recorded as well.

The goal of the study was to explore the deposition of liver glycogen after the insulin regimen in opossums. In rabbits and cats, for example, it was previously found that insulin induced significant glycogen storage.
This study found a slight depletion of liver glycogen after the insulin treatment. Our goal is to model the liver glycogen based on weight, level of blood sugar, insulin indicator, and fasting regime. Is the insulin indicator (0 no, 1 yes) an important covariate in the model?

<table>
<thead>
<tr>
<th>Animal</th>
<th>Weight</th>
<th>Liver glycogen</th>
<th>Blood sugar</th>
<th>Fasting period</th>
<th>Insulin</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1502</td>
<td>1.80</td>
<td>0.124</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1345</td>
<td>0.95</td>
<td>0.115</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1425</td>
<td>1.12</td>
<td>0.128</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1650</td>
<td>1.05</td>
<td>0.110</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1520</td>
<td>0.45</td>
<td>0.052</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1300</td>
<td>0.48</td>
<td>0.050</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1250</td>
<td>0.75</td>
<td>0.045</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1620</td>
<td>0.60</td>
<td>0.040</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1725</td>
<td>0.76</td>
<td>0.130</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1450</td>
<td>0.51</td>
<td>0.112</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1800</td>
<td>0.48</td>
<td>0.105</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1685</td>
<td>0.34</td>
<td>0.121</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>1560</td>
<td>0.38</td>
<td>0.116</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>1650</td>
<td>0.45</td>
<td>0.108</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>1650</td>
<td>0.65</td>
<td>0.032</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>1575</td>
<td>0.28</td>
<td>0.025</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1260</td>
<td>0.10</td>
<td>0.045</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>1485</td>
<td>0.26</td>
<td>0.050</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>1520</td>
<td>0.18</td>
<td>0.030</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1616</td>
<td>0.30</td>
<td>0.028</td>
<td>36</td>
<td>1</td>
</tr>
</tbody>
</table>

14.20. **Prostate Cancer Data.** This data set comes from the study by Stamey et al. (1989) that examined the relationship between the level of serum prostate specific antigen (Yang polyclonal radioimmunoassay) and a number of histological and morphometric measures in 97 patients who were about to receive a radical prostatectomy. The data are organized as data structure `prost` with first 8 fields (`prost.lcavol` - `prost.pgg45`) as predictors, and the 9th field (`prost.lpsa`) as the response.

(a) Load the data into MATLAB and run procedure stepwise. Write down the regression equation suggested by `stepwise`.

(b) Mr. Smith (a new patient) has response $y = 2.3$ and covariates:

\[ x_1 = 1.4, x_2 = 3.7, x_3 = 65, x_4 = 0.1, x_5 = 0, x_6 = -0.16, x_7 = 7, \text{ and } x_8 = 30. \]

How close to the measured response $y = 2.3$ does the regression from (a) predict $y$ for Mr. Smith? Denote this prediction by $\hat{y}_p$. Calculate the residual $r = \hat{y}_p - y$. *Hint:* In calculating $\hat{y}_p$ you should use only covariates $x_i$ suggested by `stepwise` procedure.

(c) The best, in an $R^2$ sense, single predictor for $y$ is $x_1$ – the logarithm of the cancer volume. Fit the univariate regression using $x_1$ as the predictor.
Table 14.2 Fields in structure file prost. First 8 fields are predictors, and the last is the response to be modeled.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>Logarithm of cancer volume</td>
</tr>
<tr>
<td>x2</td>
<td>Logarithm of prostate weight</td>
</tr>
<tr>
<td>x3</td>
<td>Patient’s age</td>
</tr>
<tr>
<td>x4</td>
<td>Logarithm of benign prostatic hyperplasia amount</td>
</tr>
<tr>
<td>x5</td>
<td>Seminal vesicle invasion, 0 – no, 1 – yes.</td>
</tr>
<tr>
<td>x6</td>
<td>Logarithm of capsule penetration</td>
</tr>
<tr>
<td>x7</td>
<td>Gleason score</td>
</tr>
<tr>
<td>x8</td>
<td>Percentage Gleason scores 4 or 5</td>
</tr>
<tr>
<td>y</td>
<td>Logarithm of prostate specific antigen</td>
</tr>
</tbody>
</table>

What is \( \hat{y}_p \) for Mr. Smith based on this univariate regression? Find a 95% prediction interval for \( y_p \). Is \( y_p = 2.3 \) in the interval?


Volume is the most widely used measure of wood quantity and is often estimated in standing trees for the assessment of economic value or commercial utilization potential. Volume is usually estimated from such measurements as diameter and merchantable height. The proposed equation for volume is

\[
V = a_0 D^{a_1} H^{a_2} \eta, \tag{14.9}
\]

where \( D \) is the diameter at breast height (1.3 m) and \( H \) is the merchantable height. Parameters \( a_0, a_1, \) and \( a_2 \) depend on the tree species while \( \eta \) is a multiplicative error with lognormal distribution with parameters \( \mu = 0 \) and \( \sigma^2 \).

Bruce and Schumacher (1935) provided data on 70 shortleaf pine trees consisting of \( D \) (inches), \( H \) (ft), and \( V \) (cu ft). The dataset is in file shortleaf.dat.

(a) Apply logarithmic transformation on both sides of equation (14.9) and by linear regression, estimate \( a_0, a_1, \) and \( a_2 \) and \( \sigma^2 \).

(b) For \( D = 15 \) in and \( H = 85 \) ft estimate the volume and find the 95% confidence interval for the mean response.

(c) If you are to select a single best predictor for \( \log V \), which one would you choose?

14.22. Hocking–Pendleton Data. This popular data set was constructed by Hocking and Pendleton (1982) to illustrate that an influential observation may not be outlier, and that an outlier may not be influential. The data,
given in file `hockpend.dat` is organized as a matrix of size 26 \times 4; the predictors \(x_1, x_2,\) and \(x_3\) are the first three columns, and the response \(y\) is the fourth column.

(a) Fit the linear regression model with the three covariates, report the parameter estimates and \(R^2\).

(b) Is the multicollinearity problem here?

(c) Is any of the 26 observations influential (in the sense of DFFITS, or Cook’s Distance)?

(d) Is any of the 26 observations potential outlier (in the sense of a large studentized residual)?

(e) Using forward variable selection propose a possibly simpler model.

14.23. **Squids.** Data, analyzed by Freund and Wilson (1998), were obtained on 22 squids. The dependent variable \(y\) is the weight of the squid in pounds. The predictor variables represent measurements on the beak or mouth of the squid. The data are provided in the file `squids.csv`.

<table>
<thead>
<tr>
<th>Column 1 Observation</th>
<th>(x_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column 2 Rostral length in inches</td>
<td>(x_2)</td>
</tr>
<tr>
<td>Column 3 Wing length in inches</td>
<td>(x_3)</td>
</tr>
<tr>
<td>Column 4 Rostral to notch length</td>
<td>(x_4)</td>
</tr>
<tr>
<td>Column 5 Notch to wing length</td>
<td>(x_5)</td>
</tr>
<tr>
<td>Column 6 Width in inches</td>
<td>(x_6)</td>
</tr>
<tr>
<td>Column 7 Weight of the squid in pounds</td>
<td>(y)</td>
</tr>
</tbody>
</table>

Scientists wanted to know how useful beak measurements are in predicting the weight of the squid. Answering this question was important in the study of sizes of squid eaten by sharks and tuna, since the beak is indigestible.

(a) Using multiple linear regression, estimate a linear model that expresses the squid weight \(y\) using the predictors \(x_1 - x_5\)

(b) What \(y\) is predicted for \(x_1 = 1.52, x_2 = 1.12, x_3 = 0.622, x_4 = 0.917,\) and \(x_5 = 0.324\)? What is the 95\% confidence interval for the mean response?

(c) Which of the 22 observations are influential? Answer by using either DFFITS or Cook’s distance.

(d) Is the multicollinearity a problem here? Explain.

(e) Using the forward selection procedure propose a more compact model.

(f) Redo (a) and (b) in Bayesian fashion using noninformative priors. A shell-ODC with the data is provided as `squids.odc`. Are the results of classical and Bayesian analyses similar?

14.24. **Slope in EIV Regression.** Show that the EIV regression slope in (14.7) tends to \(S_{xy}/S_{xx}\) when \(\eta \to 0\). [Hint: Apply L’Hôpital’s rule.]

14.25. **Interparticular Spacing and Wavelength in Nanoprisms 2.** In the context of Example 13.4, let \(x = (\text{separation})^{-1}\) and \(y = \log(\text{wavelength}).\)
The part of MATLAB regression output for nonoprism.dat data set is given below.

```matlab
[b, bint, r, rint, stats] = regress(y,[ones(size(x)) x])
```

```matlab
%b =
% -4.7182
% 1.6289
%
%bint =
% -5.6578 -3.7787
% 1.3061 1.9516
%
%r =
% 0.0037
% -0.0084
% ...
% 0.0058
% 0.0046
%
%rint =
% -0.0101 0.0176
% -0.0232 0.0064
% ...
% -0.0096 0.0213
% -0.0110 0.0203
%
%stats =
% 0.8545 111.5735 0.0000 0.0001
```

Using information contained in this output, answer the following questions:
(a) What is the regression equation linking \( x \) and \( y \)?
(b) Predict \( y = \log(\text{wavelength}) \) for \( x = 2.9 \). What is the wavelength for such \( x \)?
(c) What is \( R^2 \) here and how is it interpreted? What is the \( F \)-statistic here? Is it significant?
(d) The 95% confidence interval for the population slope \( \beta_1 \) is \([1.3061, 1.9516]\). Using information in this output construct a 99% confidence interval for \( \beta_1 \).

14.26. **Kodlin's Experiment Revisited.** Estimate ANCOVA parameters for Kodlin's blood pressure experiment (Example 14.8) in a Bayesian fashion using WinBUGS/OpenBUGS. Use noninformative priors. Compare classical parameter estimators from Example 14.8 with the Bayesian counterparts.
MATLAB AND WINBUGS FILES AND DATA SETS USED IN THIS CHAPTER

http://statbook.gatech.edu/Ch14.Reg/

adhesions.m, ancovafibers.m, caprolactone.m, coburnreg.m, cpeptide.m, degradation.m, diabetes.m, diagnostics.m, dissection.m, errorinvar.m, fatreg.m, fatreg1.m, fatregdiag.m, galton.m, hemo.m, histn.m, hubble.m, invitro.m, kanamycin.m, kodlin.m, myeb.m, oldfaithful.m, pedometer1.m, ratwei.m, silverzinc.m, tastechoese.m, vitalcapacity.m, vonneumann.m

ancovafibers.odc, fat.odc, mellitus.odc, regressionpred.odc, vortex.odc

adhesion.xls, alcos.dat|xls, arthritis1.dat|mat, bmp2.dat|mat|xlsx, circalbumin.dat, coburn.mat, Cpeptide.dat|mat, Cpeptideext.dat|mat, drosophila.dat|mat|xlsx, fat.dat|xlsx, galton.dat, galtoncompact.dat, kanamycin.dat, kidneyouabain.dat|mat|xlsx, Lactis.dat, nanoprism.dat, pearson.dat, pmr1.mat, prost.mat, prostate.dat, ranunculus.xlsx, shortleaf.dat, silverzinc.dat|mat, vitalcapacity.xlsx

CHAPTER REFERENCES


Chapter 15
Regression for Binary and Count Data

There are 10 types of people in the world, those who can read binary, and those who can’t.
– Anonymous

WHAT IS COVERED IN THIS CHAPTER

• Logistic Regression: Fitting and Assessing the Model
• Probit and Complementary Log-Log Links
• Poisson Regression: Fitting and Assessing the Model
• Two Case Studies: Caesarean Sections and Danish IHGA Study
• Log-linear Models in Contingency Tables

15.1 Introduction

Traditional simple or multiple linear regression assumes a normally distributed response centered at a linear combination of the predictors. For example, in simple regression, the response $y_i$ is modeled as normal $N(\beta_0 + \beta_1x_i, \sigma^2)$, where the expectation, conditional on covariate $x_i$, is a linear function of $x_i$.

For some regression scenarios this model is inadequate because the response is not normally distributed. The response could be categorical, for example, with two or more categories (“disease present–disease absent,” “survived–died,” “low–medium–high,” etc.) or be integer valued (“num-
ber with the disease,” “number of failures,” etc.), and yet a response may still depend on a covariate or a vector of covariates, \( x \). In this chapter we discuss logistic and Poisson regressions that are appropriate models for binary and counting responses.

In logistic regression, the responses are binary, coded, without loss of generality, as 0 and 1 (or as 1 and 2 in WinBUGS). In Poisson regression, the responses are nonnegative integers well modeled by a Poisson distribution in which the rate \( \lambda \) depends on one or more covariates. The covariates enter the model in a linear fashion; however, their connection with the expected response is nonlinear.

Both logistic and Poisson regressions are examples of a wide class of models called generalized linear models (GLMs). The term generalized linear model refers to models introduced by Nelder and Wedderburn (1972) and popularized by the monograph of McCullagh and Nelder (1982, second edition 1989). In a canonical GLM model, the response variable \( y_i \) is assumed to follow a distribution from the class of distributions called the exponential family, with mean \( \mu \), which is assumed to depend on covariates via their linear combination. The exponential family is a rich family of distributions and includes almost all important distributions (normal, Bernoulli, binomial, Poisson, gamma, etc.). This link between the mean \( \mu \) and covariates can be nonlinear, but the distribution of \( y_i \) depends on covariates only via their linear combination. Linear regression is a special case of GLMs for normally distributed responses, in which the mean is directly modeled by a linear combination of covariates.

### 15.2 Logistic Regression

Assume that a response \( y_i \), depending on a covariate \( x_i \), is categorical and can take two possible values. Examples of such responses include male–female, sick–healthy, alive–dead, pass–fail, success–failure, win–loss, etc. One usually assigns provisional numerical values to the responses, say 0 and 1, mainly to simplify notation. Our interest is in modeling the probability of response \( y = 1 \) given the observed covariates.

Classical least-squares regression \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \) is clearly inadequate since for unbounded \( x_i \) the linear term \( \beta_0 + \beta_1 x_i \) is unbounded as well. In addition, the residuals have only two possible values, and the variance of \( y_i \) is not free of \( x_i \).

We assume that \( y \) is Bernoulli distributed with \( \mathbb{E} y = \mathbb{P}(y = 1) = p \). Since \( \mathbb{E}(y_i | X = x_i) = \mathbb{P}(y_i = 1 | X = x_i) = p_i \) is a number between 0 and 1, it is reasonable to model \( p_i \) as \( F(\beta_0 + \beta_1 x_i) \) for some probability CDF \( F \). Equivalently,

\[
F^{-1}(p) = \beta_0 + \beta_1 x.
\]
In principle, any monotone cumulative distribution function $F$ can provide a link between the probability $p$ and the covariate(s), but the most used distributions are logistic, normal, and complementary log-log, leading to logistic, probit, and clog-log regressions. The most popular among the three is the logistic regression because its coefficients, measuring the impact of the predictors on the binary response $y$, have convenient interpretations via the log odds of the events $\{y = 1\}$. In logistic regression, $F^{-1}(p) = \log \frac{p}{1-p}$ is called the logit and denoted as $\text{logit}(p)$.

### 15.2.1 Fitting Logistic Regression

The basic statistical model for logistic regression is

$$y_i \sim \text{Ber}(p_i), \quad \text{(15.1)}$$
$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n,$$

when the responses are Bernoulli, 0 or 1.

When multiple measurements correspond to the same covariate, it is convenient to express the responses as binomial counts. That is, $n_i$ responses corresponding to covariate $x_i$ are grouped together and $y_i$ is the number of responses equal to 1:

$$y_i \sim \text{Bin}(n_i, p_i), \quad \text{(15.2)}$$
$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, k, \quad \sum_{i=1}^{k} n_i = n.$$
regression is not applicable. Estimating the model coefficients amounts to solving a nonlinear equation, and this is done by an iterative procedure. The algorithm for a single predictor is illustrated and implemented in the m-file `logisticmle.m`. There, the Newton–Raphson method is used to solve nonlinear likelihood equations and calculate coefficients $b_0$ and $b_1$ as estimators of population parameters $\beta_0$ and $\beta_1$. Details regarding the background and convergence of methods for estimating model parameters are beyond the scope of this text and can be found in McCullagh and Nelder (1989).

Once the parameters $\beta_0$ and $\beta_1$ are estimated, the probability $p_i = P(y = 1|x = x_i)$ is obtained as

$$
\hat{p}_i = \frac{\exp\{b_0 + b_1 x_i\}}{1 + \exp\{b_0 + b_1 x_i\}} = \frac{1}{1 + \exp\{-b_0 - b_1 x_i\}}.
$$

In addition to the nature of response $y_i$, there is another key difference between ordinary and logistic regressions. For linear regression, the variance does not depend on the mean $\beta_0 + \beta_1 x_i$; it is constant for all $x_i$. This is one of the assumptions for linear regression. In logistic regression, the variance is not constant; it is a function of the mean. From (15.1), $E y_i = p_i$ and $\text{Var} y_i = E y_i (1 - E y_i)$.

If $p - 1$ covariates ($p \geq 2$ parameters) are available, as is often the case, then

$$
X'_i b = \ell_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1}
$$

replaces $b_0 + b_1 x_i$, where $X'_i$ is the $i$th row of a design matrix $X$ of size $n \times p$, and $b = (b_0, b_1, \ldots, b_{p-1})'$. Now,

$$
\hat{p}_i = \frac{\exp\{X'_i b\}}{1 + \exp\{X'_i b\}} = \frac{1}{1 + \exp\{-X'_i b\}'.
$$

with $b$ maximizing the log-likelihood

$$
\log L(\beta) = \ell(\beta) = \sum_{i=1}^n y_i \cdot (X'_i \beta) - \sum_{i=1}^n \log \left(1 + \exp\{X'_i \beta\}\right).
$$

**Example 15.1. Caesarean-Section Infections.** A Caesarean-section, or C-section, is major abdominal surgery, so mothers who undergo C-sections are more likely to have an infection, excessive bleeding, blood clots, more
postpartum pain, a longer hospital stay, and a significantly longer recovery. The data in this example comes from Munich’s *Klinikum Großharden* (Fahrmeir and Tutz, 1996) and concerns infections in births by C-section. The response variable of interest is the occurrence or nonoccurrence of infection. Three covariates, each at two levels, were considered as important for the occurrence of infection:

- **noplan** – C-section delivery was planned (0) or not planned (1);
- **riskfac** – risk factors for the mother, such as diabetes, overweight, previous C-section birth, etc., are present (1) or not present (0); and
- **antibio** – antibiotics as a prophylaxis are given (1) or not given (0).

Table 15.1 provides the results:

<table>
<thead>
<tr>
<th></th>
<th>Planned</th>
<th>No plan</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Infection</td>
<td>Infection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
<td>Yes</td>
</tr>
<tr>
<td>Antibiotics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk factor yes</td>
<td>1</td>
<td>17</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>Risk factor no</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>No antibiotics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk factor yes</td>
<td>28</td>
<td>30</td>
<td>58</td>
<td>23</td>
</tr>
<tr>
<td>Risk factor no</td>
<td>8</td>
<td>32</td>
<td>40</td>
<td>0</td>
</tr>
</tbody>
</table>

Here is the MATLAB code that uses built-in functions `glmfit` and `glmval` to fit and present the model:

```matlab
infection = [1 1 1 0 0 28 23 8 0];
total = [18 98 2 0 58 26 40 9];
proportion = infection./total;
noplan = [0 1 0 1 0 1 0 1 1];
riskfac = [1 1 0 0 1 1 0 0 1];
antibio = [1 1 1 1 0 0 0 0 0];
[b,dev,stats] = glmfit([noplan' riskfac' antibio']',... [infection' total']','binomial','logit');
logitFit = ...
glmval(b,'[noplan' riskfac' antibio']','logit');
```

The resulting additive model (with no interactions) is

\[
\log \frac{P(\text{infection})}{P(\text{no infection})} = \beta_0 + \beta_1 \cdot \text{noplan} + \beta_2 \cdot \text{riskfac} + \beta_3 \cdot \text{antibio}
\]

with estimators of \( \beta \)s as

<table>
<thead>
<tr>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \beta_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.8926</td>
<td>1.0720</td>
<td>2.0299</td>
<td>-3.2544</td>
</tr>
</tbody>
</table>
The interpretation of the estimators for \( \beta \) coefficients is illuminating if we look at the odds ratio:

\[
\frac{P(\text{infection})}{P(\text{no infection})} = \exp(\beta_0) \cdot \exp(\beta_1 \text{ noplan}) \cdot \exp(\beta_2 \text{ riskfac}) \cdot \exp(\beta_3 \text{ antibio}).
\]

For example, when \( \text{antibio}=1 \), that is, when antibiotics are given, the estimated odds of infection \( \frac{P(\text{infection})}{P(\text{no infection})} \) increase by the factor \( \exp(-3.25) = 0.0388 \), that is, the odds decrease 25.79 times. Of course, these statements are valid only if the model is accurate. Other competing models, such as probit or clog-log, may result in different changes in risk ratios.

![Fig. 15.1 Caesarean delivery infection predictions. For a triple (noplan, riskfac, antibio), the numbers on the x-axis code as follows: 1 = (0, 1, 1), 2 = (1, 1, 1), 3 = (0, 0, 1), 4 = (1, 0, 1), 5 = (0, 1, 0), 6 = (1, 1, 0), 7 = (0, 0, 0), and 8 = (1, 0, 0). Blue squares are the observed relative frequencies and green circles are the model-predicted probabilities of infection. Note that point 4 does not have an observed proportion.](image-url)

The m-function `logisticmle.m` also gives standard errors for estimators of \( \beta \). Table 15.2 provides \( t \)-values, that is, ratios of coefficients and their standard deviations, for testing if the coefficients are significantly different from 0. These are known as Wald’s \( Z \) statistics, since they are approximately normal.

The deviance (page 761) of this model as a measure of goodness of fit is distributed as \( \chi^2 \) with 3 degrees of freedom. The number of degrees of freedom is calculated as 7 (the number of groups with observations) minus 4 (four estimated parameters \( \beta_0 - \beta_3 \)). Since the deviance is found significant,

\[ \text{dev} = 10.9967; \]
Table 15.2 $t$-ratios (Wald’s $Z$ statistic) for the estimators $b = \hat{\beta}$.

<table>
<thead>
<tr>
<th></th>
<th>$b$</th>
<th>$s_b$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.8926</td>
<td>0.4124</td>
<td>-4.5893</td>
</tr>
<tr>
<td>noplan</td>
<td>1.0720</td>
<td>0.4253</td>
<td>2.5203</td>
</tr>
<tr>
<td>riskfac</td>
<td>2.0299</td>
<td>0.4553</td>
<td>4.4588</td>
</tr>
<tr>
<td>antibio</td>
<td>-3.2544</td>
<td>0.4812</td>
<td>-6.7624</td>
</tr>
</tbody>
</table>

$pval = 1 - \text{chi2cdf}(\text{dev}, 7-4) \quad \% 0.0117$

the fit of this model is inadequate. To improve the fit, one may include the interactions.

One may ask why the regression model was needed in the first place. The probabilities of interest could be predicted by relative frequencies. For example, in the case $(\text{noplan} = 0, \text{riskfac} = 1, \text{antibio} = 1)$, the relative frequency of infection was $1/18 = 0.0556$, just slightly larger than the model-predicted $\hat{p} = 0.0424$. There are two benefits in using regression. First, the model is able to predict probabilities in the cases where no patients are present, such as for $(\text{noplan} = 1, \text{riskfac} = 0, \text{antibio} = 1)$. Second, the predictions for the cases where $y = 1$ is not observed are “borrowing strength” from other data and are not modeled individually. For example, zero as an estimator in the case $(\text{noplan} = 1, \text{riskfac} = 0, \text{antibio} = 0)$ is not reasonable; the model-based estimator $\hat{p} = 0.3056$ is more realistic. Figure 15.1 compares observed and model-predicted infection rates. For a triple of covariates (noplan, riskfac, antibio), the numbers on the $x$-axis code as follows: $1 = (0, 1, 1), 2 = (1, 1, 1), 3 = (0, 0, 1), 4 = (1, 0, 1), 5 = (0, 1, 0), 6 = (1, 1, 0), 7 = (0, 0, 0), \text{and } 8 = (1, 0, 0)$. Note that point 4 does not have an observed proportion; however, the model-predicted proportion can be found. For computational aspects refer to file `caesarean.m`.

Next, we provide a Bayesian solution to this example and compare the model fit with the classical fit above. The comparisons are summarized in Table 15.1.

C-SECTION INFECTIONS

```
model{
  for(i in 1:N){
    inf[i] ~ dbin(p[i], total[i])
    logit(p[i]) <- beta0 + beta1*noplan[i] +
    beta2*riskfac[i] + beta3*antibio[i]
  }
  beta0 ~dnorm(0, 0.00001)
  beta1 ~dnorm(0, 0.00001)
  beta2 ~dnorm(0, 0.00001)
  beta3 ~dnorm(0, 0.00001)
}
```

DATA
list(inf=c(1, 11, 0, 0, 28, 23, 8, 0),
       total = c(18, 98, 2, 0, 58, 26, 40, 9),
       noplan = c(0,1,0,1,0,1,0,1),
       riskfac = c(1,1, 0, 0, 1,1, 0, 0),
       antibio =c(1,1,1,0,0,0,0,0), N=8)

INITS

list(beta0=0, beta1=0, beta2=0, beta3=0)

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>val2.5pc</th>
<th>median</th>
<th>val97.5pc</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>-1.964</td>
<td>0.4258</td>
<td>0.001468</td>
<td>-2.853</td>
<td>-1.945</td>
<td>-1.183</td>
<td>1001</td>
<td>1000000</td>
</tr>
<tr>
<td>beta1</td>
<td>1.111</td>
<td>0.4339</td>
<td>8.857E-4</td>
<td>0.2851</td>
<td>1.102</td>
<td>1.986</td>
<td>1001</td>
<td>1000000</td>
</tr>
<tr>
<td>beta2</td>
<td>2.104</td>
<td>0.4681</td>
<td>0.00159</td>
<td>1.226</td>
<td>2.09</td>
<td>3.066</td>
<td>1001</td>
<td>1000000</td>
</tr>
<tr>
<td>beta3</td>
<td>-3.335</td>
<td>0.4915</td>
<td>9.756E-4</td>
<td>-4.337</td>
<td>-3.322</td>
<td>-2.411</td>
<td>1001</td>
<td>1000000</td>
</tr>
<tr>
<td>deviance</td>
<td>32.24</td>
<td>2.873</td>
<td>0.00566</td>
<td>28.67</td>
<td>31.59</td>
<td>39.49</td>
<td>1001</td>
<td>1000000</td>
</tr>
</tbody>
</table>

Table 15.3 Comparison of classical and noninformative Bayes estimators $b = \hat{b}$, with estimators of standard deviations.

15.2.2 Assessing the Logistic Regression Fit

The measures for assessing the goodness of linear regression fit that we covered in Chapter 14, $R^2$, $F$, MSE, etc., are not appropriate for logistic regression. As in the case of linear regression, there is a range of measures for assessing the performance of logistic regression, and we will briefly discuss a few.

The significance of model parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ is tested by the so-called Wald’s test. One finds the statistic $Z_i = \frac{b_i}{s_i(b_i)}$ that has an approximate normal distribution if the coefficient $\beta_i$ is 0. Equivalently, $W_i = \frac{b_i^2}{s^2(b_i)}$ with an approximate $\chi^2$-distribution with 1 degree of freedom can be used. Large values of $|Z_i|$ or $W_i$ are critical for $H_0 : \beta_i = 0$.

The sample variances $s^2(b_i)$ are diagonal elements of $(X'VX)^{-1}$, where
15.2 Logistic Regression 761

\[ X = \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\
1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\
\vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1}
\end{bmatrix} \]

is the design matrix and

\[ V = \begin{bmatrix}
\hat{p}_1(1 - \hat{p}_1) & 0 & \cdots & 0 \\
0 & \hat{p}_2(1 - \hat{p}_2) & \cdots & 0 \\
\vdots \\
0 & 0 & \cdots & \hat{p}_n(1 - \hat{p}_n)
\end{bmatrix}. \]

The customary measure for goodness of fit is **deviance**, defined as

\[ D = -2\log \frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}. \]

For the logistic regression in (15.2), where \( y_i \) is the number of 1s and \( n_i - y_i \) is the number of 0s in class \( i \), the likelihood is

\[ L = \prod_{i=1}^{k} p_i^{y_i}(1 - p_i)^{n_i-y_i} \]

and the deviance is

\[ D = -2 \sum_{i=1}^{k} \left\{ y_i \log \left( \frac{\hat{y}_i}{y_i} \right) + (n_i - y_i) \log \left( \frac{n_i - \hat{y}_i}{n_i - y_i} \right) \right\}, \]

where \( \hat{y}_i = n_i \hat{p}_i \) is the model fit for \( y_i \). The saturated model estimates \( p_i \) as \( \hat{p}_i = y_i/n_i \) and \( \hat{y}_i = y_i \), providing the fit that matches the observations.

The deviance statistic in this case has a \( \chi^2 \)-distribution with \( k - p \) degrees of freedom, where \( k \) is the number of classes/groups and \( p \) is the number of parameters in the model. Recall that in the previous example the deviance of the model was distributed as \( \chi^2 \) with \( k - p = 7 - 4 = 3 \) degrees of freedom.

For both Bernoulli and binomial observations, the mean and variance depend on a single parameter, \( p \). When the mean is well fitted, the variance may be underfitted (overdispersion in data) or overfitted (underdispersion in data). The ratio \( D/df \) is often used to indicate over- or underdispersion in the data.

The traditional \( \chi^2 \)-statistic for the goodness of fit in model (15.2) is defined as

\[ \chi^2 = \sum_{i=1}^{k} \left[ \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} + \frac{(y_i - \hat{y}_i)^2}{n_i - \hat{y}_i} \right], \]

where \( n_i \) is the number of observations in class \( i, i = 1, \ldots, k \). This statistic has an approx. \( \chi^2 \)-distribution with \( k - p \) degrees of freedom.
Goodness of fit measure $G$ is defined as the difference of deviance between the null model (intercept-only model) and the model under consideration. $G$ has a $\chi^2$-distribution with $p - 1$ degrees of freedom, and small values of $G$ are critical, suggesting that the deviance did not improve significantly by adding covariates.

The logistic model can always be expressed in terms of Bernoulli outcomes, where $y_i$ is 0 or 1, as in (15.1). Then $k = n$, $n_i = 1$, the likelihood for the saturated model, is $\prod_{i=1}^{n} y_i^{y_i} (1 - y_i)^{1 - y_i} = 1$ (we assume that $0^0 = 1$), and the deviance for the Bernoulli representation becomes

$$D = -2 \sum_{i=1}^{n} [y_i \log \hat{p}_i + (1 - y_i) \log (1 - \hat{p}_i)].$$

Statistic $D$ does not follow any specific distribution, regardless of the sample size. Likewise, the Pearson $\chi^2$ becomes

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)}$$

(15.3)
in model (15.1) and does not follow any specific distribution, either.

To further evaluate the model, several types of residuals are available. Deviance residuals are defined as

$$r^D_i = \text{sign}(y_i - \hat{y}_i) \sqrt{2 \left\{ y_i \log \left( \frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \log \left( \frac{n_i - y_i}{n_i - \hat{y}_i} \right) \right\}}, \quad i = 1, \ldots, k,$

for model (15.2) and

$$r^D_i = \text{sign}(y_i - \hat{p}_i) \sqrt{2y_i \log \hat{p}_i + (1 - y_i) \log (1 - \hat{p}_i),} \quad i = 1, \ldots, n,$

for model (15.1). The deviance $D$ is decomposed to the sum of squares of deviance residuals in an ANOVA-like fashion as $D = \sum (r^D_i)^2$. The squared residual $(r^D_i)^2$ measures the contribution of the $i$th case to the deviance.

Deviance residuals can be plotted against the order of sampling to explore for possible trends and outliers. Also useful for checking the model are half-normal plots where the ordered absolute values $r^D_i$ are plotted against the normal quantiles $\Phi^{-1}\left(\frac{i + n + 1/8}{2n + 1/2}\right)$. These kinds of plots are an extension of Atkinson’s (1985) half-normal plots in regular linear regression models. Deviation from a straight line in a half-normal plot indicates model inadequacy.

For the model in (15.1), the Pearson residual is defined as

$$r^\text{pea}_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}},$$
and the sum of squares of $r_i^{pea}$ constitutes Pearson’s $\chi^2$ statistic,

$$\sum_{i=1}^{n} (r_i^{pea})^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)},$$

as in (15.3). This statistic represents a discrepancy measure; however, as we mentioned, it does not follow the $\chi^2$-distribution, even asymptotically.

In the case of continuous covariates, large $n$ and small $n_i$, Hosmer and Lemeshow proposed a $\chi^2$-statistic based on the grouping of predicted values $\hat{p}_i$. All $\hat{p}_i$ are ordered and divided into $g$ approximately equal groups, usually 10. For 10 groups, sample deciles of ordered $\hat{p}_i$ can be used.

The Hosmer–Lemeshow statistic is

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{(n_i - n\bar{p}_i)^2}{n\bar{p}_i},$$

where $g$ is the number of groups, $n_i$ is the number of cases in the $i$th group, and $\bar{p}_i$ is the average of model (predicted) probabilities for the cases in the $i$th group. The $\chi^2_{HL}$ statistic is compared to $\chi^2_{g-2}$ quantiles, and small $p$-values indicate that the fit is poor. In the case of ties, that is, when there are blocks of items with the same predicted probability $\hat{p}$, the blocks are not split but assigned to one of the two groups that share the block. The details of the algorithm can be found in Hosmer and Lemeshow (1989).

In the case of linear regression, $R^2$, as a proportion of model-explained variability in observations, has a strong intuitive appeal in assessing the regression fit. In the case of logistic regression, there is no such intuitive $R^2$. However, there are several proposals of $R^2$-like measures, called pseudo-$R^2$. Most of them are defined in terms of model likelihood or log-likelihood. The model likelihood and log-likelihood are calculated using the logit model,

$$\ell_i = b_0 + b_1x_{i1} + \cdots + b_{p-1}x_{i,p-1},$$

$$LL_p = LL(b_0, \ldots, b_{p-1}) = \sum_{i=1}^{n} (y_i \times \ell_i - \log(1 + \exp(\ell_i))),$$

and the model likelihood is $L_p = \exp\{LL_p\}$. The null model is fitted without covariates, and

$$\ell_0 = b_0,$$

$$LL_{null} = LL(b_0) = \sum_{i=1}^{n} (y_i \times \ell_0 - \log(1 + \exp(\ell_0))).$$

The null model likelihood is $L_{null} = \exp\{LL_{null}\}$.

By analogy to linear regression, $R^2 = \frac{SSR}{SST} = \frac{SSR}{SST}$. 

\[ R^2_{mf} = \frac{LL_{null} - LL_p}{LL_{null}} = 1 - \frac{LL_p}{LL_{null}}, \]

defines McFadden’s pseudo-\( R^2 \). Some other counterparts of \( R^2 \) are

- **Cox–Snell**: \( R^2_{cs} = 1 - \left[ \frac{L_{null}}{L_p} \right]^{2/n}; \)
- **Nagelkerke**: \( R^2_n = \frac{1 - \left[ \frac{L_{null}}{L_p} \right]^{2/n}}{1 - \left[ \frac{L_{null}}{L_{null}} \right]^{2/n}}; \)
- **Efron**: \( R^2_e = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}. \)

**Example 15.2. Arrhythmia.** Patients who undergo coronary artery bypass graft (CABG) surgery have an approximately 19% to 40% chance of developing atrial fibrillation (AF). AF is a quivering, chaotic motion in the upper chambers of the heart, known as the atria. AF can lead to the formation of blood clots, causing greater in-hospital mortality, strokes, and longer hospital stays. While this can be alleviated with drugs, it is very expensive and sometimes dangerous if not warranted. Ideally, several risk factors that would indicate an increased risk of developing AF in this population could save lives and money by indicating which patients need pharmacological intervention. Researchers began collecting data from CABG patients during their hospital stay such as demographics like age and sex, as well as heart rate, cholesterol, operation time, etc. Then the researchers recorded which patients developed AF during their hospital stay. The goal was to evaluate the probability of AF given the measured demographic and risk factors.

The data set `arrhythmia.dat`, courtesy of Dr. Matthew C. Wiggins, contains the following variables:

| \( Y \) | Fibrillation |
| \( X_1 \) | Age |
| \( X_2 \) | Aortic cross clamp time |
| \( X_3 \) | Cardiopulmonary bypass time |
| \( X_4 \) | Intensive care unit (ICU) time |
| \( X_5 \) | Average heart rate |
| \( X_6 \) | Left ventricle ejection fraction |
| \( X_7 \) | Anamnesis of hypertension |
| \( X_8 \) | Gender [1 - female; 0 - male] |
| \( X_9 \) | Anamnesis of diabetes |
| \( X_{10} \) | Previous MI |

The MATLAB script `arrhythmia.m` provides a logistic regression fit. The script calculates deviance and several goodness-of-fit measures.
15.2 Logistic Regression

load 'Arrhythmia.mat'
Y = Arrhythmia(:,1);
X = Arrhythmia(:,2:11);  %Design matrix n x (p-1) without
vector 1 (intercept)
Xdes = [ones(size(Y)) X]; %with the intercept: n x p
n = length(Y); %number of subjects
alpha = 0.05; %alpha for CIs
[b, dev, stats]=glmfit(X,Y, 'binomial','link','logit')
lin = Xdes * b  %linear predictor, n x 1 vector

Figure 15.2 shows observed arrhythmia responses (0 or 1) with their logistic fit.

![Arrhythmia responses 0 or 1 with their logistic fit. The abscise axis is the linear predictor lin.](image)

With the linear predictor, fitted probabilities for \( Y_i = 1 \) are given as \( \hat{p}_i \). The estimators of the \( \beta \)s with their standard deviations and \( p \)-values for the Wald test are given next. The intercept is significantly nonzero (0.0158), and the variable \( x_1 \) (age) is strongly significant (0.0005). This agrees with the inference based on confidence intervals; only the intervals for \( \beta_0 \) and \( \beta_1 \) do not contain 0, or, equivalently, the intervals for the odds ratio, \( \exp(\beta_0) \) and \( \exp(\beta_1) \), do not contain 1.

phat = exp(lin)./(1 + exp(lin));
V = diag( phat .* (1 - phat) );
sqrtV = diag( sqrt(phat .* (1 - phat) ) )
sb = sqrt( diag( inv( Xdes’ * V * Xdes ) ) )
% inv( Xdes’ * V * Xdes') is stats.covb
% Wald tests for parameters beta
z = b./sb %tests for beta.i = 0, i=0,...,p-1
pvals = 2 * normcdf(abs(z)) %p-values
%[0.0158; 0.0005; 0.3007; 0.2803; 0.1347; 0.8061
% 0.4217; 0.3810; 0.6762; 0.0842; 0.5942]
%(1-alpha)*100% CI for betas
CIs = [b - norminv(1 - alpha/2) * sb , b + norminv(1-alpha/2) * sb]
%(1-alpha)*100% CIs for odds ratios
CIsOR = exp([b-norminv(1-alpha/2)*sb , b+norminv(1-alpha/2)+sb])
% 0.0000 0.1281
% 1.0697 1.2711
% 0.9781 1.0744
% ...
% 0.8628 10.3273
% 0.4004 4.9453

Figure 15.3 shows estimators of $\beta_0 - \beta_{10}$ (as green circles) and 95% confidence bounds. Since the intervals for $\beta_0$ and $\beta_1$ do not contain 0, both the intercept and covariate age are important in the model. It is tempting to do variable/model selection based on outcomes of Wald’s test— but this is not advisable. Exclusion of a parameter/variable from the model will necessarily change the estimators and confidence intervals for the remaining parameters and previously insignificant parameters may become significant. As in linear regression, best subset, forward, and backward variable selection procedures exist and may be implemented.

Fig. 15.3 Estimators of $\beta_0 - \beta_{10}$ are shown as green circles, and 95% confidence intervals are given. For comparison, the intervals for $\beta_1 - \beta_6$ are shown separately on a different scale.
Next, we find the log-likelihoods for the model and null model. The model deviance is 78.2515, while the difference of deviances between the models is 26.1949. This would be a basis for a likelihood ratio test if the response were grouped. Since in a Bernoulli setup the distributions of deviance and $G$ are not $\chi^2$, the testing needs to be done by one of the response-grouping methods, such as the Hosmer–Lemeshow method.

\[
\loglik = \sum (Y \cdot \text{lin} - \log(1 + \exp(\text{lin}))) \quad -39.1258
\]

% fitting null model.
\[
[b0, dev0, stats0] = \text{glmfit(zeros(size(Y)), Y, 'binomial', 'link', 'logit')}
\]
\[
\% b0=-0.6381, dev0=104.4464, stats=... (structure)
\]
\[
\loglik0 = \sum (Y \cdot b0(1) - \log(1 + \exp(b0(1)))) \quad -52.2232
\]
\[
G = -2 \cdot (\loglik0 - \loglik) \quad 26.1949
\]
\[
dev0 - \text{dev} \quad 26.1949, \text{ the same as } G, \text{ difference of deviances}
\]
\[
\% \text{model deviance}
\]
\[
devi = -2 \cdot \sum (Y \cdot \log(\text{phat} + \text{eps}) + (1-Y) \cdot \log(1 - \text{phat} + \text{eps}))
\]
\[
\% 78.2515, \text{ directly}
\]
\[
dev = -2 \cdot \loglik \quad 78.2515, \text{ glmfit output}
\]
\[
\% -2 \cdot \loglik \quad 78.2515, \text{ as a link between } \loglik \text{ and deviance}
\]

Several measures correspond to $R^2$ in the linear regression context: McFadden’s pseudo-$R^2$, Cox–Snell $R^2$, Nagelkerke $R^2$, and Efron’s $R^2$. All measures fall between 0.25 and 0.4.

\[
\% \text{McFadden Pseudo } R^2, \text{ equivalent expressions}
\]
\[
mcfadden = \frac{-2 \cdot (\loglik0 - \loglik)}{-2 \cdot \loglik0} \quad 0.2508
\]
\[
1 - \frac{\loglik}{\loglik0} \quad 0.2508
\]
\[
% \text{coxsnell} = 1 - (\exp(\loglik0)/\exp(\loglik))^{(2/n)} \quad 0.2763
\]
\[
nagelkerke = (1 - (\exp(\loglik0)/\exp(\loglik))^{(2/n)})/...
\]
\[
(1 - \exp(\loglik0)^{(2/n)}) \quad 0.3813
\]
\[
% \text{effron} = 1 - \sum((Y - \text{phat})^2)/\sum((Y - \text{sum}(Y)/n)^2) \quad 0.2811
\]

Next we find several types of residuals: ordinary, Pearson, deviance, and Anscombe.

\[
ro = Y - \text{phat}; \quad \% \text{Ordinary residuals}
\]
\[
% \text{Deviance Residuals}
\]
\[
rdev = \text{sign}(Y - \text{phat}) \cdot \sqrt{-2 \cdot Y \cdot \log(\text{phat} + \text{eps})} - ... \\
\]
\[
2 \cdot (1 - Y) \cdot \log(1 - \text{phat} + \text{eps}));
\]
\[
% \text{Anscombe Residuals}
\]
\[
ransc = \text{betainc}(Y, 2/3, 2/3) - \text{betainc}(\text{phat}, 2/3, 2/3) \quad ... \\
\]
\[
(\text{phat} \cdot \log(1 - \text{phat} + \text{eps})^{(1/6)};
\]
\[
% \model deviance is recovered as
\]
\[
% \text{the sum of squared dev. residuals}
\]
\[
\text{sum(rdev}^2) \quad 78.2515
\]
Figure 15.4 shows four kinds of residuals (ordinary, Pearson, deviance, and Anscombe), plotted against $\hat{p}$.

![Graphs showing four types of residuals plotted against $\hat{p}$](image)

Fig. 15.4 Ordinary, Pearson, deviance, and Anscombe residuals plotted against $\hat{p}$.

If the model is adequate, the smoothed residuals should result in a function close to 0. Figure 15.5 shows Pearson’s residuals smoothed by a loess smoothing method (loess.m).

Influential and outlying observations can be detected with a plot of absolute values of residuals against half-normal quantiles. Figure 15.6 was produced by the script below and shows a half-normal plot. The upper and lower bounds (in red) show an empirical 95% confidence interval and were obtained by simulation. The sample of size 19 was obtained from Bernoulli $\text{Ber}(\hat{p})$, where $\hat{p}$ is the model fit, and then the minimum, mean, and maximum of the absolute residuals of the simulated values were plotted.

```matlab
k = 1:n;
q = norminv((k + n - 1/8)./(2 * n + 1/2));
plot( q, sort(abs(rdev)), 'k-','LineWidth',1.5);

% Simulated Envelope
rand('state',1)
env =[];
for i = 1:19
    surrogate = binornd(1, phat);
    rdevsu = sign(surrogate - phat).*sqrt(- 2*surrogate .* ...
                   log(phat+eps)- 2*(1 - surrogate) .* log(1 - phat+eps) );
    env = [env sort(abs(rdevsu))];
```

768 15 Regression for Binary and Count Data
To predict the mean response for a new observation, we selected a "new person" with specific covariate values. For this person the estimator for $P(Y = 1)$ is 0.3179, and 0 for a single future response. A single future response is in fact a classification problem: individuals with a specific set of covariates are classified as either 0 or 1.

% Probability of $Y=1$ for a new observation
Xh = [1 72 81 130 15 78 43 1 0 0 1]';
% responses for a new person
pXh = exp(Xh' * b) / (1 + exp(Xh' * b)) %0.3179
%(1-alpha) * 100% CI
ppXh = Xh' * b %-0.7633
s2pXp = Xh' * inv( Xdes' * V * Xdes ) * Xh %0.5115
spXh = sqrt(s2pXp) %0.7152
% confidence interval on the linear part
li = [ppXh-norminv(1-alpha/2)*spXh ...
ppXh+norminv(1-alpha/2)*spXh] %-2.1651 0.6385
% transformation to the CI for the mean response
exp(li)./(1 + exp(li)) %0.1029 0.6544

Fig. 15.5 Pearson's residuals (green circles) smoothed. The red circles show the result of smoothing.
% Predicting single future observation

cutoff = sum(Y)/n \times 0.3457

pXh > cutoff \Rightarrow Y_{new} = 0

Next, we provide a Bayesian solution to the Arrhythmia logistic model (Arrhythmia.odc) and compare classical and Bayesian model parameters.

model{
  eps <- 0.00001
  for(i in 1:N){
    Y[i] ~ dbern(p[i])
    devres[i] <- 2*Y[i] * log(Y[i]/p[i] + eps) +
      2*(1 - Y[i]) * log((1-Y[i])/(1-p[i]) + eps)
  }
  for(j in 1:11){
    beta[j] ~ dnorm(0, 0.0001)
  }
  dev <- sum(devres[])
}

DATA + INITS (see Arrhythmia.odc)

The classical and Bayesian model parameters are shown in the table:
15.2 Logistic Regression

$$\hat{\beta}_0 \hat{\beta}_1 \hat{\beta}_2 \hat{\beta}_3 \hat{\beta}_4 \hat{\beta}_5 \hat{\beta}_6 \hat{\beta}_7 \hat{\beta}_8 \hat{\beta}_9 \hat{\beta}_{10}$$

Deviation

<table>
<thead>
<tr>
<th></th>
<th>Classical</th>
<th>Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_0$</td>
<td>-10.95</td>
<td>-13.15</td>
</tr>
<tr>
<td>$\hat{\beta}_1$</td>
<td>0.1536</td>
<td>0.1863</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>0.0248</td>
<td>0.0335</td>
</tr>
<tr>
<td>$\hat{\beta}_3$</td>
<td>-0.1295</td>
<td>-0.1541</td>
</tr>
<tr>
<td>$\hat{\beta}_4$</td>
<td>0.0207</td>
<td>0.0106</td>
</tr>
<tr>
<td>$\hat{\beta}_5$</td>
<td>-0.5377</td>
<td>-0.6419</td>
</tr>
<tr>
<td>$\hat{\beta}_6$</td>
<td>0.3416</td>
<td>0.4027</td>
</tr>
<tr>
<td>$\hat{\beta}_7$</td>
<td>78.25</td>
<td>89.89</td>
</tr>
<tr>
<td>$\hat{\beta}_8$</td>
<td>1.936</td>
<td>1.313</td>
</tr>
<tr>
<td>$\hat{\beta}_9$</td>
<td>0.3416</td>
<td>0.4027</td>
</tr>
<tr>
<td>$\hat{\beta}_{10}$</td>
<td>78.25</td>
<td>89.89</td>
</tr>
</tbody>
</table>

15.2.3 Probit and Complementary Log-Log Links

We have seen that for logistic regression,

$$\hat{p}_i = F(\ell_i) = \frac{\exp\{\ell_i\}}{1 + \exp\{\ell_i\}},$$

where $\ell_i = b_0 + b_1 x_{i1} + \cdots + b_{p-1} x_{i,p-1}$ is the linear part of the model.

A probit regression uses a normal distribution instead,

$$\hat{p}_i = \Phi(\ell_i),$$

while for the complementary log-log, the extreme value (Gumbel type I for the minimum) distribution

$$F(x) = 1 - \exp\{-\exp\{x\}\}$$

is used.

The complementary log-log link interprets the regression coefficients in terms of the hazard ratio rather than the log odds ratio. It is defined as

$$\text{clog-log} = \log(-\log(1 - p)).$$

The clog-log regression is typically used when the outcome $\{y = 1\}$ is rare. See also Remark on page 775 on a link between clog-log and Poisson regressions. Unlike the logit and probit links, the clog-log link is asymmetric, that is, $\text{clog-log}(p) \neq -\text{clog-log}(1 - p)$.

Probit models are popular in a bioassay context. A disadvantage of probit models is that the link $\Phi^{-1}$ does not have an explicit expression, although approximations and numerical algorithms for its calculation are readily available.

Once the linear part $\ell_i$ in a probit or clog-log model is fitted, the probabilities are estimated as

$$\hat{p}_i = \Phi(\ell_i) \quad \text{or} \quad \hat{p}_i = 1 - \exp(-\exp(\ell_i)),$$

respectively. In MATLAB, the probit and complementary log-log links are optional arguments, `'link','probit'` or `'link','comploglog'`.

Example 15.3. Bliss Data. In his 1935 paper, Bliss provides a table showing a number of flour beetles killed after 5 hours of exposure to gaseous car-
bon disulfide at various concentrations. This data set has since been used extensively by statisticians to illustrate and compare models for binary and binomial data.

Table 15.4 Bliss beetle data.

<table>
<thead>
<tr>
<th>Dose (log$_{10}$CS$_2$mg$^{-1}$)</th>
<th>Number of Beetles</th>
<th>Number Killed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6907</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td>1.7242</td>
<td>60</td>
<td>13</td>
</tr>
<tr>
<td>1.7552</td>
<td>62</td>
<td>18</td>
</tr>
<tr>
<td>1.7842</td>
<td>56</td>
<td>28</td>
</tr>
<tr>
<td>1.8113</td>
<td>63</td>
<td>52</td>
</tr>
<tr>
<td>1.8369</td>
<td>59</td>
<td>53</td>
</tr>
<tr>
<td>1.8610</td>
<td>62</td>
<td>61</td>
</tr>
<tr>
<td>1.8839</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

The following Bayesian model is applied on the Bliss data, and a probit fit is provided (bliss.odc).

```r
model{
  for (i in 1:N) {
    y[i] ~ dbin(p[i], n[i])
    probit(p[i]) <- alpha.star + beta * (x[i] - mean(x[]))
    yhat[i] <- n[i] * p[i]
  }
  alpha <- alpha.star - beta * mean(x[])
  beta ~ dnorm(0, 0.001)
  alpha.star ~ dnorm(0, 0.001)
}
```

**DATA**

```r
list(x = c(1.6907, 1.7242, 1.7552, 1.7842,
          1.8113, 1.8369, 1.8610, 1.8839),
     n = c(59, 60, 62, 56, 63, 59, 62, 60),
     y = c(6, 13, 18, 28, 52, 53, 61, 60), N = 8)
```

**INITS**

```r
list(alpha.star=0, beta=0)
```

```
mean   sd    MC error  val2.5pc median val97.5pc start sample
alpha  -35.03 2.652 0.01837  -40.35  -35.01 -29.98  1001 100000
alpha.star  0.4461 0.07724 0.00070  0.2938  0.4461  0.5973  1001 100000
beta    19.78 1.491 0.01040  16.94   19.77  22.78  1001 100000
yhat[1]  3.445 1.018 0.00683  1.757   3.336  5.725  1001 100000
yhat[2]  10.76 1.69 0.009674  7.643   10.74 14.26  1001 100000
yhat[3]  23.48 1.896 0.01095  19.77   23.47 27.2  1001 100000
yhat[4]  33.81 1.597 0.01072  30.62   33.83 36.85  1001 100000
yhat[5]  49.59 1.623 0.01208  46.28   49.63 52.64  1001 100000
yhat[6]  53.26 1.158 0.008777  50.8    53.33 55.33  1001 100000
yhat[7]  59.59 0.7477 0.00561  57.91   59.68 60.82  1001 100000
yhat[8]  59.17 0.3694 0.00271  58.28   59.23 59.71  1001 100000
```
If instead of probit, the clog-log was used, as $\text{cloglog}(p[x]) <- \alpha^{\star} + \beta(x[x] - \text{mean}(x[x]))$, then the coefficients are

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>MC error 2.5pc</th>
<th>median</th>
<th>val97.5pc</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>-39.73</td>
<td>3.216</td>
<td>-46.24</td>
<td>-39.66</td>
<td>-33.61</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>$\beta$</td>
<td>22.13</td>
<td>1.786</td>
<td>18.73</td>
<td>22.09</td>
<td>25.74</td>
<td>1001</td>
<td>100000</td>
</tr>
</tbody>
</table>

For comparisons we take a look at the classical solution (beetleBliss2.m). Figure 15.7 shows three binary regressions (logit, probit and clog-log) fitting the Bliss data.

The table below compares the coefficients of the linear part of the three models. Note that classical and Bayesian results are close because the priors in the Bayesian model are noninformative.
15.3 Poisson Regression

Poisson regression models the counts $y = \{0, 1, 2, 3, \ldots\}$ of rare events in a large number of trials. Typical examples are unusual adverse events, accidents, incidence of a rare disease, device failures during a particular time interval, etc. Recall that Poisson random variable $Y \sim \text{Poi}(\lambda)$ has the probability mass function

$$f(y) = \mathbb{P}(Y = y) = \frac{\lambda^y}{y!} \exp\{-\lambda\}, \quad y = 0, 1, 2, 3, \ldots,$$

with both mean and variance equal to the rate parameter $\lambda > 0$.

Suppose that $n$ counts of $y_i, i = 1, \ldots, n$ are observed and that each count corresponds to a particular value of a covariate $x_i, i = 1, \ldots, n$. A typical Poisson regression can be formulated as follows:

$$y_i \sim \text{Poi}(\lambda_i),$$

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n,$$

although other relations between $\lambda_i$ and the linear part $\beta_0 + \beta_1 x_i$ are possible as long as $\lambda_i$ remains positive. More generally, $\lambda_i$ can be linked to a linear expression containing $p - 1$ covariates and $p$ parameters as

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}, i = 1, \ldots, n.$$
In terms of model (15.4) the Poisson rate $\lambda_i$ is the expectation, and its logarithm can be expressed as $\log \mathbb{E}(y_i|X = x_i) = \beta_0 + \beta_1 x_i$. When the covariate $x_i$ gets a unit increment, $x_i + 1$, then

$$\log \mathbb{E}(y_i|X = x_i + 1) = \beta_0 + \beta_1 x_i + \beta_1 = \log \mathbb{E}(y_i|X = x_i) + \beta_1.$$ 

Thus, parameter $\beta_1$ can be interpreted as the increment to log rate when the covariate gets an increment of 1. Equivalently, $\exp\{\beta_1\}$ is the ratio of rates,

$$\exp\{\beta_1\} = \frac{\mathbb{E}(y_i|x_i + 1)}{\mathbb{E}(y_i|x_i)}.$$ 

The model-assessed mean response is $\hat{y}_i = \exp\{b_0 + b_1 x_i\}$, where $b_0$ and $b_1$ are the estimators of $\beta_0$ and $\beta_1$. Strictly speaking, the model predicts the rate $\hat{\lambda}_i$, but the rate is interpreted as the expected response.

**Remark.** If a Poisson regression model

$$y_i \sim \text{Poi}(\lambda_i), \quad \lambda_i = \exp\{b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}\} = \exp\{\ell_i\}, \quad i = 1, \ldots, n,$$

is dichotomized as

$$y^*_i = \begin{cases} 1, & y_i > 0 \\ 0, & y_i = 0 \end{cases},$$

then an adequate model for $y^*_i$ is a binary regression with the clog-log link. Indeed,

$$\mathbb{P}(y^*_i = 1) = 1 - \mathbb{P}(y^*_i = 0) = 1 - \exp\{-\lambda_i\} = 1 - \exp\{-\exp(\ell_i)\}.$$

The deviance of the model, $D$, is defined as

$$D = 2 \sum_{i=1}^{n} \left( y_i \log \frac{y_i}{\hat{y}_i} - (y_i - \hat{y}_i) \right),$$

where $y_i \log y_i = 0$ if $y_i = 0$. As in logistic regression, the deviance is a measure of goodness of fit of a model and for a Poisson model has a $\chi^2$-distribution with $n - p$ degrees of freedom.

Deviance residuals, defined as

$$r^\text{dev}_i = \text{sign}(y_i - \hat{y}_i) \times \sqrt{2y_i \log \frac{y_i}{\hat{y}_i} - 2(y_i - \hat{y}_i)},$$
satisfy \( D = \sum_{i=1}^{n} \left( r_i^{dev} \right)^2 \). Note that the sum of squares of the deviance residuals simplifies to \( D = 2\sum_{i=1}^{n} y_i \log(y_i/\hat{y}_i) \), since in the Poisson regression \( \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0 \).

Pearson’s residuals are defined as

\[
 r_i^{pea} = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}.
\]

Then the Pearson goodness-of-model-fit statistic \( \chi^2 = \sum_{i=1}^{n} (r_i^{pea})^2 \) also has a \( \chi^2 \)-distribution with \( n - p \) degrees of freedom. Although homoscedastic, those residuals are asymmetric.

Freedman–Tukey residuals are defined as

\[
 r_i^{ft} = \sqrt{y_i} + \sqrt{y_i + 1} - \sqrt{4\hat{y}_i + 1}
\]

and Anscombe residuals (closest to normality) as

\[
 r_i^a = \frac{3}{2} \left( \frac{y_i^{2/3} - \hat{y}_i^{2/3}}{\hat{y}_i^{1/6}} \right).
\]

The pseudo-\( R^2 \) for the Poisson regression is defined as \( 1 - D / D_0 \). Here \( D \) is deviance of the model in question, and \( D_0 \) is deviance of the intercept-only model,

\[
 D_0 = 2 \sum_{i=1}^{n} y_i \log(y_i/\bar{y}),
\]

where \( \bar{y} \) is the sample mean of the observations.

Some additional diagnostic tools are exemplified in the following case study (\texttt{ihga.m}):

\textbf{Example 15.4. Case Study: Danish IHGA Data.} In an experiment conducted in the 1980s (Hendriksen et al., 1984), 572 elderly people living in a number of villages in Denmark were randomized, 287 to a control (C) group who received standard care, and 285 to an experimental group who received standard care plus IHGA: a kind of preventive assessment in which each person’s medical and social needs were assessed and acted upon individually. The important outcome was the number of hospitalizations during the 3-year life of the study.

\[
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
 \%
### Table 15.5 Distribution of number of hospitalizations in IHGA study.

<table>
<thead>
<tr>
<th>Group</th>
<th># of hospitalizations</th>
<th>n</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>138 77 46 12 8 4 0 2 287</td>
<td>0.944</td>
<td>1.54</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>147 83 37 13 3 1 1 0 285</td>
<td>0.768</td>
<td>1.02</td>
<td></td>
</tr>
</tbody>
</table>

```matlab
3*ones(12,1); 4 * ones(8,1); 5*ones(4,1); 7*ones(2,1)];
y1 = [0*ones(147,1); 1*ones(83,1); 2*ones(37,1);...
3*ones(13,1); 4 * ones(3,1); 5*ones(1,1); 6*ones(1,1)];

%response # of hospitalizations
x = [x0; x1]; y = [y0; y1];
xdes = [ones(size(y)) x];
[n p] = size(xdes)
[b dev stats] = glmfit(x,y,'poisson','link','log')
yhat = glmval(b, x,'log') %model predicted responses

% Pearson residuals
rpea = (y - yhat)./sqrt(yhat);
% deviance residuals
rdev = sign(y - yhat) .* sqrt(-2*y.*log(yhat./(y + eps)) - 2*(y - yhat));
% Friedman-Tukey residuals
rft = sqrt(y) + sqrt(y + 1) - sqrt(4*yhat + 1)
% Anscombe residuals
ransc = 3/2 * (y.^(2/3) - yhat.^(2/3))./(yhat.^(1/6))

Figure 15.8 shows four our types of residuals in Poisson regression fit of IHGA data: Pearson, deviance, Friedman–Tukey, and Anscombe. The residuals are plotted against responses y.

loglik = sum(y .* log(yhat + eps) - yhat - log(factorial(y)));
% [b0, dev0, stats0] = glmfit(zeros(size(y)),y,'poisson','link','log')
yhat0 = glmval(b0, zeros(size(y)),'log');
loglik0 = sum(y .* log(yhat0 + eps) - yhat0 - log(factorial(y)));
G = - 2 * (loglik0 - loglik) %LR test, nested model chi2 5.1711
dev0 - dev % the same as G, difference of deviances 5.1711
pval = 1-chi2cdf(G,1) %0.0230

Under $H_0$ stating that the model is null (model with an intercept and no covariates), the statistic $G$ will have $df = p - 1$ degrees of freedom, in our case $df = 1$. Since this test is significant ($p = 0.0230$), the covariate contributes significantly to the model.

Below are several ways to express the deviance of the model.

%log-likelihood for saturated model
logliksat = sum(y .* log(y+eps) - y - log(factorial(y))) %-338.1663
m2LL = -2 * sum( y .* log(yhat./(y + eps)) ) %819.8369
deviance = sum(rdev.^2) %819.8369
dev %819.8369 from glmfit
-2*(loglik - logliksat) % 819.8369
Fig. 15.8 (a) Four types of residuals in a Poisson regression fit of IHGA data: Pearson, deviance, Friedman–Tukey, and Anscombe. The residuals are plotted against responses $y$.

The following is a Bayesian model fit in WinBUGS (geriatric.odc).
Example 15.5. Cellular Differentiation Data. In a biomedical study of the immunoactivating ability of the agents TNF (tumor necrosis factor) and IFN
(interferon) to induce cell differentiation, the number of cells that exhibited markers of differentiation after exposure to TNF or IFN was recorded (Piergorsch et al., 1988; Fahrmeir and Tutz, 1994). At each of the 16 dose combinations of TNF/IFN, 200 cells were examined. The number \( y \) of differentiating cells corresponding to a TNF/IFN combination are given in Table 15.6.

<table>
<thead>
<tr>
<th>Number of cells diff</th>
<th>Dose of TNF (U/ml)</th>
<th>Dose of IFN (U/ml)</th>
<th>Number of cells diff</th>
<th>Dose of TNF (U/ml)</th>
<th>Dose of IFN (U/ml)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>31</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>4</td>
<td>68</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>20</td>
<td>69</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>39</td>
<td>0</td>
<td>100</td>
<td>128</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0</td>
<td>102</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>38</td>
<td>1</td>
<td>4</td>
<td>171</td>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>52</td>
<td>1</td>
<td>20</td>
<td>180</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>69</td>
<td>1</td>
<td>100</td>
<td>193</td>
<td>100</td>
<td>20</td>
</tr>
</tbody>
</table>

The suggested model is Poisson with the form

\[
\lambda = \mathbb{E}(y|\text{TNF, IFN}) = \exp\{\beta_0 + \beta_1 \text{TNF} + \beta_2 \text{IFN} + \beta_3 \text{TNF} \times \text{IFN}\}.
\]

From

```matlab
load 'celular.dat'
number = celular(:,1);
TNF = celular(:,2);
IFN = celular(:,3);
[b, dev, stats] = glmfit([TNF IFN TNF.*IFN],number,...
    'poisson','link','log')
```

the estimators for \( \beta_0-\beta_3 \) are

\[
\begin{align*}
\beta_0 & = 3.43463 \\
\beta_1 & = 0.01553 \\
\beta_2 & = 0.00895 \\
\beta_3 & = -0.0000567
\end{align*}
\]

Since \( \beta_3 < 0.0001 \), it is tempting to drop the interaction term. However, since the standard error of \( \beta_3 \) is \( \text{s.e.}(\beta_3) = 0.000013484 \), Wald’s \( Z = \beta_3/\text{s.e.}(\beta_3) \) statistic is \( -4.2050 \), suggesting that the term \( \text{TNF} \times \text{IFN} \) might be significant. However, the overdispersion parameter, which theoretically should be \( \text{stats.s}=1 \), is estimated as \( \text{stats.s}=3.42566 \), and the Wald statistic should be adjusted to \( Z' = -0.0000567/(3.42566 \times 0.000013484) = 1.2275 \). Since the \( p \)-value is \( 2 \times \text{normcdf}(-1.2275) = 0.2196 \), after all, the interaction term turns out not to be significant and the additive model could be fit:

```matlab
[b, dev, stats] = glmfit([TNF IFN],number,...
    'poisson','link','log')
```
which gives estimates $b_0 = 3.57311$, $b_1 = 0.01314$, and $b_2 = 0.00585$. Details can be found in \texttt{celular.m}.

The additive model was also fit in a Bayesian manner.

```plaintext
model {
  for (i in 1:n) {
    numbercells[i] ~ dpois(lambda[i])
    lambda[i] <- exp(beta0 + beta1 * tnf[i] + beta2 * ifn[i])
  }
  beta0 ~ dnorm(0, 0.00001)
  beta1 ~ dnorm(0, 0.00001)
  beta2 ~ dnorm(0, 0.00001)
}

DATA
list(n=16,
  numbercells = c(11,18,20,39,22,38,52,69,31,68,69,128,102,171,180,193),
  tnf = c(0,0,0, 1,1,1, 10,10,10, 100,100,100,100,100),
  ifn = c(0,4,20,100, 0,2,20,100, 0,4,20,100, 0,4,20,100 ) )

INITS
list(b0=0, b1=0, b2=0)
```

Note that, because of the noninformative priors on $\beta_0$, $\beta_1$, and $\beta_2$, the Bayesian estimators almost coincide with the MLEs from \texttt{glmfit}.

15.4 Log-linear Models

Poisson regression can be employed in the context of contingency tables (Chapter 12). The logarithm of the table cell count is modeled in a linear fashion.

In an $r \times c$ contingency table, let the probability of a cell $(i,j)$ be $p_{ij}$. If $n$ subjects are cross-tabulated, let $n_{ij}$ be the number of subjects classified in the cell $(i,j)$. The count $n_{ij}$ is realization of random variable $N_{ij}$. We assume that the sample size $n$ is random, since in that case the cell frequency $N_{ij}$ is a Poisson random variable with the intensity $\mu_{ij}$. If the sample size $n$ is fixed, then the $N_{ij}$s are realizations of a multinomial random variable. In Fisher’s
exact test context (page 602), we saw that if in addition the marginal counts are fixed, the $N_{ij}$s are hypergeometric random variables.

Then the expected table is

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>$\cdots$</th>
<th>c</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$np_{11}$</td>
<td>$np_{12}$</td>
<td>$np_{1c}$</td>
<td>$np_1$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$np_{21}$</td>
<td>$np_{22}$</td>
<td>$np_{2c}$</td>
<td>$np_2$</td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>$np_{r1}$</td>
<td>$np_{r2}$</td>
<td>$np_{rc}$</td>
<td>$np_r$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$np_1$</td>
<td>$np_2$</td>
<td>$np_c$</td>
<td>$n$</td>
<td></td>
</tr>
</tbody>
</table>

Note that both $n_{ij}$ and $e_{ij} = n_i \times n_j / n$ involve observations and empirical marginal probabilities defined by the observed marginal frequencies. If we denote $\mu_{ij} = \mathbb{E}N_{ij} = np_{ij}$, then both $n_{ij}$ and $e_{ij}$ are estimators of $\mu_{ij}$; the first is unconstrained and the second is constrained by the assumption of independence of factors, $p_{ij} = p_i \cdot p_j$. This point is important since $e_{ij}$ are “expected” under independence once the table is observed and fixed, while $\mu_{ij}$ are expectations of the random variables $N_{ij}$.

The log-linear model for the expected frequency $\mu_{ij}$ is given as

$$\log \mu_{ij} = \lambda_0 + \lambda_i^R + \lambda_j^C + (\lambda^{RC})_{ij}, \quad i = 1, \ldots, r; j = 1, \ldots, c,$$

where $\lambda_i^R$ and $\lambda_j^C$ are contributions by row and column, respectively, and $\lambda_{ij}^{RC}$ is a row–column interaction term.

This model is called a saturated log-linear model and is similar to the two-factor ANOVA model. Note that an unconstrained model has $(1 + r + c + r \times c)$ free parameters but only $r \times c$ observations, $n_{ij}$. To make this over-parameterized model identifiable, constraints on parameters are imposed. A standard choice is STZ: $\sum_i \lambda_i^R = 0$, $\sum_j \lambda_j^C = 0$, $\sum_i (\lambda^{RC})_{ij} = \sum_j (\lambda^{RC})_{ij} = 0$. Thus, with the constraints we have $1 - (r - 1) + (c - 1) + (r - 1)(c - 1) = r \times c$ free parameters, which is equal to the number of observations, and the model gives a perfect fit for the observed frequencies. This is the reason why this model is called saturated.

The hypothesis of independence in a contingency table that was discussed Chapter 12 has simple form:

$$H_0 : \lambda_{ij}^{RC} = 0, \quad i = 2, \ldots, r; j = 2, \ldots, c.$$  

Under $H_0$ the log-linear model becomes additive:

$$\log \mu_{ij} = \lambda_0 + \lambda_i^R + \lambda_j^C, \quad i = 1, \ldots, r; j = 1, \ldots, c.$$  

The MLEs of components in the log-linear model (not derived here, but see Agresti, 2002) are
\[ \hat{\lambda}_0 = \frac{\sum_{i,j} \log n_{ij}}{rc}, \]
\[ \hat{\lambda}_i^R = \frac{\sum_j \log n_{ij}}{c} - \hat{\lambda}_0, \]
\[ \hat{\lambda}_j^C = \frac{\sum_i \log n_{ij}}{r} - \hat{\lambda}_0, \]
\[ \hat{\lambda}_{ij}^{RC} = \log n_{ij} - (\hat{\lambda}_0 + \hat{\lambda}_i^R + \hat{\lambda}_j^C). \]

If any \( n_{ij} \) is equal to 0, then all entries in the table are replaced by \( n_{ij} + 0.5 \).

Traditional analysis involves testing that particular \( \lambda \) components are equal to 0. One approach, often implemented in statistical software, would be to find the variance of \( \hat{\lambda} \), \( \text{Var}(\hat{\lambda}) \), using a hypergeometric model, and then use the statistic \( (\hat{\lambda})^2 / \text{Var}(\hat{\lambda}) \) that has a \( \chi^2 \)-distribution with one degree of freedom (\( \lambda \) here is any of \( \lambda_0, \lambda_i^R, \lambda_j^C \), or \( \lambda_{ij}^{RC} \)). Large values of this statistic are critical for \( H_0 \).

Next, we focus on the Bayesian analysis of a log-linear model.

**Example 15.6. Log-linear Model and Bystanders.** In the psychological experiment of Exercise 12.9, seeking assistance (help) was dependent on a subject’s perception of the number of bystanders. The resulting chi-square statistic \( \chi^2 = 7.908 \) was significant with a \( p \)-value of about 2%. We revisit this exercise and provide a Bayesian solution using a log-linear approach. The WinBUGS program `bystanders.odc` is used to conduct statistical inference.

```plaintext
model{
  for (i in 1:r) for (j in 1:c)  n[i,j] ~ dpois(mu[i,j]);
  log(mu[i,j]) <- lambda0 + lambdaR[i] + lambdaC[j] + lambdaRC[i,j]
}

lambda0 ~ dnorm(0,prec)

lambdaR[1] <- 0
lambdaC[1] <- 0
for (i in 2:r) { lambdaR[i] ~ dnorm(0,prec) }
for (i in 2:c) { lambdaC[i] ~ dnorm(0,prec) }

for (j in 1 : c) { lambdaRC[1, j] <- 0 }
for (i in 2 : r) { lambdaRC[i, 1] <- 0;
  for (j in 2 : c) { lambdaRC[i, j] ~ dnorm(0, prec) }
}

DATA
list(r=3,c=2,prec=0.0001,
n=structure(.Data=c(11,2,16,10,4,9),.Dim=c(3,2)))
The hypothesis of independence is assessed by testing that all \( \lambda_{RC} \) are equal to 0. In this case \( \lambda_{RC[1,1]} \), \( \lambda_{RC[1,2]} \), and \( \lambda_{RC[2,1]} \) are set to 0 because of identifiability constraints. The 95% credible set for interaction \( \lambda_{RC[2,2]} \) contains 0 – therefore, this interaction is not significant. However, \( \lambda_{RC[3,2]} \) is significantly positive since the credible set \([0.9045, 5.041]\) does not contain 0.

Example 15.7. **Upton–Fingleton Square.** An example from Upton and Fingleton (1985) concerns finding directional trends in spatial count data. The simple count data set is given as follows:

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

We are interested in testing if there are any north–south or east–west trends present in the spatial pattern. The idea of Upton and Fingleton was to establish a Poisson regression where the response was the intensity \( n_{ij} \) in the location \((i, j)\) and the covariates are \( x \)-coordinate, related to the east–west trend, and \( y \)-coordinate, related to the north–south trend.

Thus, the observed frequencies \( n_{ij} \) will be modeled as

\[
E(n_{ij}) = \exp(\beta_0 + \beta_1 x_i + \beta_2 y_j), \quad i, j = 1, \ldots, 5,
\]

where \( x_i = i \) and \( y_j = j \). A significant \( \beta_1 \) or \( \beta_2 \) in a well-fitting Poisson regression will indicate the presence of corresponding trends (see **UptonFingleton.m** for details).
15.5 Exercises

15.1. Blood Pressure and Heart Disease. This example is based on data given by Cornfield (1962). A sample of male residents of Framingham, Massachusetts, aged 40–59, were chosen. During a 6-year follow-up period they were classified according to several factors, including blood pressure and whether they had developed coronary heart disease. The results for these two variables are given in the table below. The covariate for blood pressure represents an interval, for example, 122 stands for blood pressure in the interval 117–126.

<table>
<thead>
<tr>
<th>Blood pressure</th>
<th>No disease</th>
<th>Disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>112 (&lt;117)</td>
<td>156</td>
<td>3</td>
<td>159</td>
</tr>
<tr>
<td>122 (117–126)</td>
<td>252</td>
<td>17</td>
<td>269</td>
</tr>
<tr>
<td>132 (127–136)</td>
<td>284</td>
<td>12</td>
<td>296</td>
</tr>
<tr>
<td>142 (137–146)</td>
<td>271</td>
<td>16</td>
<td>287</td>
</tr>
<tr>
<td>152 (147–156)</td>
<td>139</td>
<td>12</td>
<td>151</td>
</tr>
<tr>
<td>162 (157–166)</td>
<td>85</td>
<td>8</td>
<td>93</td>
</tr>
<tr>
<td>177 (167–186)</td>
<td>99</td>
<td>16</td>
<td>115</td>
</tr>
<tr>
<td>197 (&gt;186)</td>
<td>43</td>
<td>8</td>
<td>51</td>
</tr>
</tbody>
</table>

Using logistic regression, estimate the probability of disease for a person with an average blood pressure equal to 158.

15.2. Blood Pressure and Heart Disease in WinBUGS. Use data from the previous exercise. Parameter \( p \) represents the probability of developing coronary disease and can be estimated as

\[
\hat{p} = \frac{e^{b_0 + b_1 BP}}{1 + e^{b_0 + b_1 BP}},
\]

where \( b_0 \) and \( b_1 \) are Bayes estimators of \( \beta_0 \) and \( \beta_1 \) obtained by WinBUGS. Use noninformative priors on \( \beta_0 \) and \( \beta_1 \).

(a) What are the values \( b_0 \) and \( b_1 \)? Provide plots of posterior densities for \( b_0 \) and \( b_1 \).

(b) What are the 95% credible intervals for \( \beta_0 \) and \( \beta_1 \)?

(c) Estimate the probability of disease for a person with an average blood pressure of 158.

15.3. Sex of Diamond-backed Terrapins and Incubation Temperature. Temperature-dependent sex determination, observed in some reptiles and fish, is a type of environmental sex determination in which the temperatures experienced during embryonic development determine the sex of the offspring. Below are data on the relationship between the ratio of male/female diamond-backed terrapins (Malaclemys terrapin) and incubation temperature, as reported by Burke and Calichio (2014):
Predict the probability of a female terrapin for a temperature of 29°C.

15.4. **Health Promotion.** Students at the University of the Best in England (UBE) investigated the use of a health promotion video in a doctor’s surgery. Covariates **Age** and **Amount of Weekly Exercise** for a sample of 30 men were obtained, and each man was asked a series of questions on the video. On the basis of the responses to these questions the psychologist simply recorded whether the promotion video was **Effective** or **Not effective**. The collected data are provided in Table 15.7.

<table>
<thead>
<tr>
<th>Number</th>
<th>Age</th>
<th>Exercise</th>
<th>Code of response</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>3</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>5</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>10</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>4</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>3</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
<td>14</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>7</td>
<td>41</td>
<td>5</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>8</td>
<td>27</td>
<td>5</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>9</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>10</td>
<td>34</td>
<td>2</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>11</td>
<td>51</td>
<td>8</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>12</td>
<td>33</td>
<td>21</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>13</td>
<td>23</td>
<td>12</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>14</td>
<td>41</td>
<td>19</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>15</td>
<td>38</td>
<td>2</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>16</td>
<td>25</td>
<td>9</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>17</td>
<td>40</td>
<td>6</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>18</td>
<td>36</td>
<td>25</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>19</td>
<td>40</td>
<td>13</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>20</td>
<td>39</td>
<td>3</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>21</td>
<td>45</td>
<td>10</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>22</td>
<td>39</td>
<td>5</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>23</td>
<td>40</td>
<td>2</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>24</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>25</td>
<td>47</td>
<td>9</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>26</td>
<td>31</td>
<td>17</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>27</td>
<td>37</td>
<td>7</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>28</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>Not effective</td>
</tr>
<tr>
<td>29</td>
<td>32</td>
<td>13</td>
<td>1</td>
<td>Effective</td>
</tr>
<tr>
<td>30</td>
<td>25</td>
<td>10</td>
<td>0</td>
<td>Not effective</td>
</tr>
</tbody>
</table>

Table 15.7 Age, exercise, and effectiveness of the video.
(a) Use the fitted model to estimate the probability that a male of age 40 who exercises 10 hours a week would find the video “effective.”

Comment: The linear predictor is

\[-7.498 + 0.17465 \text{Age} + 0.16324 \text{Exercise}.\]

The positive coefficient with covariate Age means that older subjects tend to respond “video is effective” with higher probabilities. Similarly, the positive coefficient with predictor Exercise indicates that increasing the values of Exercise also increases the probabilities for the response “video is effective.”

(b) Comment on the fit of the model based on deviance and residuals.

15.5. IOP. Laser refractive surgery often decreases Intraocular Pressure (IOP) and may lead to hypotony (clinically significant low IOP that may lead to corneal decompensation, accelerated cataract formation, maculopathy, and discomfort).

An investigator wished to determine whether the post-operative IOP in patients after laser refractive surgery was related to the residual thickness of the cornea. In a sample of 140 patients who had undergone laser surgery, post-operative IOP and the thickness of the cornea were measured. The data set is provided in iop2.dat which consists of two columns, (1) indicator of low IOP (IOP < 10) and (2) central corneal thickness (in micrometers).

(a) Fit the logistic regression with cornea thickness as the predictor of incidence of low IOP.

(b) For a person who had a refractive surgery with residual thickness of cornea of 480 micrometers, what is the risk of a low IOP?

(c) Compare deviances for two links: logit (as in (a)), probit and comploglog. Which link provides the best fit? (Hint: Deviances are in the output of glmfit).

15.6. PONV. Despite advances over the past decade, including the advent of 5-HT3 receptor antagonists, combination therapy, and multimodal strategies, postoperative nausea and vomiting (PONV) remains a serious and frequent adverse event associated with surgery and anesthesia. PONV can be very distressing for patients, can lead to medical complications, and impose economic burdens. A meta-analysis of several studies gives rates of 37% for nausea and 20% for vomiting in patients undergoing general anesthesia. However, indiscriminate prophylaxis is not recommended (the “prevent-or-cure” dilemma).

There are considerable variations in the reported incidence of PONV, which can be attributed to a number of factors. Risk factors for PONV can be divided into patient risk factors, procedural risk factors, anesthetic risk factors, and postoperative risk factors. The main and well-
understood risk factors are gender, history of motion sickness/PONV, smoking status, and use of postoperative opioids.

A data set \texttt{PONV.xls} or \texttt{PONV.mat} (courtesy of Jelena Velickovic, MD anesthesiologist from Belgrade) contains records of 916 patients consisting of some demographic, anamnetic, clinical, and procedural variables. Several variables are of interest to be modeled and controlled: manifestation of PONV from 0 to 2 hours after surgery, PONV from 2 to 24 hours, and PONV from 0 to 24 hours (\texttt{PONV0to2}, \texttt{PONV2to24}, and \texttt{PONV0to24}).

Three score variables (SinclairScore, ApfelScore, and LelaScore) summarize the relevant demographic and clinical information prior to surgery with the goal of predicting PONV.

The starter file \texttt{ponv.m} gives a basic explanation of variables and helps to read all the variables into MATLAB.

Fit a logistic model for predicting \texttt{PONV0to24} based on a modified Sinclair Score defined as \( \text{MSS} = \text{SinclairScore} + \frac{1}{20} \times \text{LelaScore}. \) What is the probability of \texttt{PONV0to24} for a person with \( \text{MSS} = 1.3 \)?

15.7. **Mannose-6-phosphate Isomerase.** McDonald (1985) counted allele frequencies at the mannose-6-phosphate isomerase (Mpi) locus in the amphipod crustacean \textit{Megalorchestia californiana}, which lives on sandy beaches of the Pacific coast of North America. There were two common alleles, Mpi90 and Mpi100. The latitude of each collection location, the number of each allele, and the proportion of the Mpi100 allele are shown here:

<table>
<thead>
<tr>
<th>Location</th>
<th>Latitude</th>
<th>Mpi90</th>
<th>Mpi100</th>
<th>Prop Mpi100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Port Townsend, WA</td>
<td>48.1</td>
<td>47</td>
<td>139</td>
<td>0.748</td>
</tr>
<tr>
<td>Neskowin, OR</td>
<td>45.2</td>
<td>177</td>
<td>241</td>
<td>0.577</td>
</tr>
<tr>
<td>Siuslaw River, OR</td>
<td>44.0</td>
<td>1087</td>
<td>1183</td>
<td>0.521</td>
</tr>
<tr>
<td>Umpqua River, OR</td>
<td>43.7</td>
<td>187</td>
<td>175</td>
<td>0.483</td>
</tr>
<tr>
<td>Coos Bay, OR</td>
<td>43.5</td>
<td>397</td>
<td>671</td>
<td>0.628</td>
</tr>
<tr>
<td>San Francisco, CA</td>
<td>37.8</td>
<td>40</td>
<td>14</td>
<td>0.259</td>
</tr>
<tr>
<td>Carmel, CA</td>
<td>36.6</td>
<td>39</td>
<td>17</td>
<td>0.304</td>
</tr>
<tr>
<td>Santa Barbara, CA</td>
<td>34.3</td>
<td>30</td>
<td>0</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Estimate the probability of seeing allele Mpi100, taking into account latitude as a covariate, using logistic regression.

15.8. **Arthritis Treatment Data.** The data were obtained from Koch and Edwards (1988) for a double-blind clinical trial investigating a new treatment for rheumatoid arthritis. In this data set, there were 84 subjects of different ages who received an active or placebo treatment for their arthritis pain, and the subsequent extent of improvement was recorded as marked, some, or none. The dependent variable \texttt{improve} was an ordinal categorical observation with three categories (0 for none, 1 for some, and 2 for marked). The three explanatory variables were \texttt{treatment} (1 for active or 0 for placebo), \texttt{gender} (1 for male, 2 for female), and \texttt{age}
(recorded as a continuous variable). The data in arthritis2.dat is organized as a matrix, with 84 rows corresponding to subjects and 5 columns containing ID number, treatment, gender, age, and improvement status:

<table>
<thead>
<tr>
<th>ID</th>
<th>Treatment</th>
<th>Gender</th>
<th>Age</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>1</td>
<td>1</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>90</td>
<td>1</td>
<td>1</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>46</td>
<td>1</td>
<td>1</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>2</td>
<td>66</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>74</td>
<td>2</td>
</tr>
<tr>
<td>71</td>
<td>0</td>
<td>2</td>
<td>68</td>
<td>1</td>
</tr>
</tbody>
</table>

Dichotomize the variable improve as improve01 = improve > 0; and fit the binary regression with improve01 as a response and treatment, gender, and age as covariates. Use the three links logit, probit, and comploglog and compare the models by comparing the deviances.

15.9. **Third-degree Burns.** The data for this exercise, discussed in Fan et al. (1995), refer to \( n = 435 \) adults who were treated for third-degree burns by the University of Southern California General Hospital Burn Center. The patients were grouped according to the area of third-degree burns on the body. For each midpoint of the groupings \( \log(\text{area} + 1) \), the number of patients in the corresponding group who survived and the number who died from the burns was recorded:

<table>
<thead>
<tr>
<th>Log(area+1)</th>
<th>Survived</th>
<th>Died</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.35</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>1.60</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>1.75</td>
<td>67</td>
<td>2</td>
</tr>
<tr>
<td>1.85</td>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>1.95</td>
<td>71</td>
<td>8</td>
</tr>
<tr>
<td>2.05</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>2.15</td>
<td>35</td>
<td>31</td>
</tr>
<tr>
<td>2.25</td>
<td>7</td>
<td>49</td>
</tr>
<tr>
<td>2.35</td>
<td>1</td>
<td>12</td>
</tr>
</tbody>
</table>

(a) Fit the logistic regression on the probability of death due to third-degree burns with the covariate \( \log(\text{area} + 1) \).
(b) Using your model, estimate the probability of survival for a person for which \( \log(\text{area} + 1) \) equals 2.

15.10. **Diabetes Data.** The data repository of Andrews and Herzberg (1985) features a data set containing measures of blood glucose, insulin levels, relative weights, and clinical classifications of 145 subjects in diabetes.dat:
The columns represent the following variables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>relwt</td>
<td>Relative weight</td>
</tr>
<tr>
<td>glufast</td>
<td>Fasting plasma glucose</td>
</tr>
<tr>
<td>glutest</td>
<td>Test plasma glucose</td>
</tr>
<tr>
<td>instest</td>
<td>Plasma insulin during test</td>
</tr>
<tr>
<td>sspg</td>
<td>Steady-state plasma glucose</td>
</tr>
<tr>
<td>group</td>
<td>Clinical group: (3) overt diabetic; (2) chem. diabetic; (1) normal</td>
</tr>
</tbody>
</table>

From the variable group form the variable dia and set it to 1 if diabetes is present (group=1, 2) and 0 if the subject has no diabetes (group=1). Find the regression of the five other variables on dia.

15.11. **Remission Ratios over Time.** A clinical trial on a new anticancer agent produced the following remission ratios for 40 patients on trial at each of the six stages of the trial:

<table>
<thead>
<tr>
<th>Stage</th>
<th>Remission Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9/40</td>
</tr>
<tr>
<td>2</td>
<td>14/40</td>
</tr>
<tr>
<td>3</td>
<td>22/40</td>
</tr>
<tr>
<td>4</td>
<td>29/40</td>
</tr>
<tr>
<td>5</td>
<td>33/40</td>
</tr>
<tr>
<td>6</td>
<td>35/40</td>
</tr>
</tbody>
</table>

Fit the logistic model for the probability of remission if the stages are measured in equal time units. Give the probability that a new patient on this regiment will be in remission at stage 4 and discuss how this probability compares to 29/40.

15.12. **Death of Sprayed Flour Beetles.** Hewlett (1974) and Morgan (2000) provide data that have been considered by many researchers in bioassay theory. The data consist of a quantal bioassay for pyrethrum in which the mortality of adult flour beetles (Tribolium castaneum) was measured over time under four dose levels. The columns are cumulative numbers of dead adult flour beetles exposed initially to pyrethrum, a well-known plant-based insecticide. Mixed with oil, the pyrethrum was sprayed at the given dosages over small experimental areas in which the groups of beetles were confined but allowed to move freely. The beetles were fed during the experiment in order to eliminate the effect of natural mortality.
Using WinBUGS, model the proportion of dead beetles using sex, dosage, and day as covariates. The data set in WinBUGS format can be found in `tribolium.odc`.

15.13. **Mortality in Swiss White Mice.** An experiment concerning the influence of diet on the rate of *Salmonella enteritidis* infection (mouse typhoid) among W-Swiss mice was conducted by Schneider and Webster (1945). Two diets, one of whole wheat and whole dried milk (coded 100) and the other synthetic (coded 191), are compared against two doses of bacilli, 50K and 500K. The experimental results are summarized in the following table.

<table>
<thead>
<tr>
<th>Diet</th>
<th>Dose 50K</th>
<th>Dose 500K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of mice surviving</td>
<td>Number of mice surviving</td>
</tr>
<tr>
<td>100</td>
<td>293</td>
<td>144</td>
</tr>
<tr>
<td>191</td>
<td>296</td>
<td>141</td>
</tr>
</tbody>
</table>

The authors conclude that diet is able to condition natural resistance, but one of the factors was the genetic constitution of the mice employed. Model the probability of survival using an additive logistic regression with the covariates Diet and Dose.

15.14. **Kyphosis Data.** The measurements in `kyphosis.dat` are from Hastie and Tibshirani (1990, p. 301) and were collected on 83 patients undergoing corrective spinal surgery (Bell et al., 1994). The objective was to determine the important risk factors for kyphosis, or the forward flexion of the spine at least 40 degrees from vertical following surgery. The covariates are age in months, the vertebrae level at which the surgery started, and the number of vertebrae involved.
The data set kyphosis.dat has four columns:

<table>
<thead>
<tr>
<th>Age</th>
<th>Start</th>
<th>Number</th>
<th>Kyphosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>158</td>
<td>14</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>128</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>13</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>6</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>13</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

where the first three columns are age, start, and number, and where the fourth column is binary indicator for the presence of kyphosis, 0 if absent and 1 if present.

Using logistic regression, model the probability of present kyphosis given the risk factors age, start, and number.

15.15. **Prostate Cancer.** The prostate cancer clinical trial data of Byar and Green (1980) is given in the file prostatecanc.dat with a description of the variables in prostatecanc.txt or prostatecanc.m. There are 475 observations with 12 measured covariates. The response is the stage (3 or 4) of a patient assessed by a physician.

Propose a model for predicting the stage by a subset of predictors.

15.16. **Pediculosis Capitis.** An outbreak of *Pediculosis capitis* is being investigated in a girls’ school containing 291 pupils. Of 130 children who live in a nearby housing estate, 18 were infested, and of 161 who live elsewhere, 37 were infested. Thus, the school girls are stratified by the housing attribute into two groups: (A) the nearby housing estate and (B) elsewhere.

(a) Test the hypothesis that the population proportions of infested girls for groups A and B are the same.

(b) Run a logistic regression that predicts the probability of a girl being infested including the predictor housing that takes value 1 if the girl is from group A and 0 if she is from group B. All you need are the sample sizes from groups A and B and the corresponding incidences of infestation. You might need to recode the data and represent the summarized data as 291 individual cases containing the incidence of infestation and housing status.

(c) A sample of 26 girls from group A and 34 from group B are randomly selected for more detailed modeling analysis. The instances of infestation (0-no, 1-yes), housing (A=1, B=0), family income (in thousands), family size, and girl’s age are recorded. The data are available in the file lice.xls. Propose the logistic model that predicts the probability that a girl who is infested will possess some or all of the predictors (housing, income, size, age). This is an open-ended question, and you are expected to defend your proposed model.
(c.1) According to your model, what is the probability of a girl being infested if housing is A, family income is 74, family size is 4, and age is 12? Of course, you will use only the values of the predictors included in your model.

(c.2) If the family size is increased by 1 and all other covariates remain the same, how much do the odds of infestation change?

(d) The 55 affected girls were divided randomly into two groups of 29 and 26. The first group received a standard local application and the second group a new local application. The efficacy of each was measured by clearance of the infestation after one application. By this measure the standard application failed in ten cases and the new application in five. Is the new treatment more effective? This part may mimic the methodology in (a).

15.17. **Finney Data.** In a controlled experiment to study the effect of the rate and volume of air inspired on a transient reflex vasoconstriction in the skin of the digits, 39 tests under various combinations of rate and volume of air inspired were conducted (Finney, 1947). The end point of each test was whether or not vasoconstriction occurred.

<table>
<thead>
<tr>
<th>Volume Rate</th>
<th>Response</th>
<th>Volume Rate</th>
<th>Response</th>
<th>Volume Rate</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.70 0.825</td>
<td>Constrict</td>
<td>3.50 1.09</td>
<td>Constrict</td>
<td>1.80 1.50</td>
<td>Constrict</td>
</tr>
<tr>
<td>1.25 2.50</td>
<td>Constrict</td>
<td>0.75 1.50</td>
<td>Constrict</td>
<td>0.95 1.90</td>
<td>No constrict</td>
</tr>
<tr>
<td>0.80 3.20</td>
<td>Constrict</td>
<td>0.70 3.50</td>
<td>Constrict</td>
<td>1.90 0.95</td>
<td>Constrict</td>
</tr>
<tr>
<td>0.60 0.75</td>
<td>No constrict</td>
<td>1.10 1.70</td>
<td>No constrict</td>
<td>1.60 0.40</td>
<td>No constrict</td>
</tr>
<tr>
<td>0.90 0.75</td>
<td>No constrict</td>
<td>0.90 0.45</td>
<td>No constrict</td>
<td>2.70 0.75</td>
<td>Constrict</td>
</tr>
<tr>
<td>0.80 0.57</td>
<td>No constrict</td>
<td>0.55 2.75</td>
<td>No constrict</td>
<td>2.35 0.03</td>
<td>No constrict</td>
</tr>
<tr>
<td>0.60 3.00</td>
<td>No constrict</td>
<td>1.40 2.33</td>
<td>Constrict</td>
<td>1.10 1.83</td>
<td>No constrict</td>
</tr>
<tr>
<td>0.75 3.75</td>
<td>Constrict</td>
<td>2.30 1.64</td>
<td>Constrict</td>
<td>1.10 2.20</td>
<td>Constrict</td>
</tr>
<tr>
<td>3.20 1.60</td>
<td>Constrict</td>
<td>0.85 1.415</td>
<td>Constrict</td>
<td>1.20 2.00</td>
<td>Constrict</td>
</tr>
<tr>
<td>1.70 1.06</td>
<td>No constrict</td>
<td>1.80 1.80</td>
<td>Constrict</td>
<td>0.80 3.33</td>
<td>Constrict</td>
</tr>
<tr>
<td>0.40 2.00</td>
<td>No constrict</td>
<td>0.95 1.36</td>
<td>No constrict</td>
<td>0.95 1.90</td>
<td>No constrict</td>
</tr>
<tr>
<td>1.35 1.35</td>
<td>No constrict</td>
<td>1.50 1.36</td>
<td>No constrict</td>
<td>0.75 1.90</td>
<td>No constrict</td>
</tr>
<tr>
<td>1.60 1.78</td>
<td>Constrict</td>
<td>0.60 1.50</td>
<td>No constrict</td>
<td>1.30 1.625</td>
<td>Constrict</td>
</tr>
</tbody>
</table>

Model the probability of vasoconstriction as a function of two covariates, Volume and Rates. Use MATLAB and compare the results with WinBUGS output. The data in MATLAB and WinBUGS formats is provided in finney.dat.

15.18. **Diagnosing Sagittal Synostosis.** (Courtesy of Dr. Marcus Walker.) In early human development, the skull is made up of many different bones, and the gaps between these bones are called cranial sutures. During the first few years after birth, these cranial sutures allow the bones to grow and the skull to expand to make room for the growing brain. These bones naturally grow closer together and fuse to form one solid skull for protection, but if any of these cranial sutures fuse too early while the brain is still rapidly growing, a condition known as craniosynostosis occurs that results in skull deformation and constriction of the developing brain. Craniosynostosis occurs in approximately 1 in every 2,000 chil-
Children can lead to developmental disabilities, blindness, and even death if left untreated. One specific type of craniosynostosis, called sagittal synostosis, is characterized by a premature fusion of the suture that runs from the front to the back of the skull. This has been observed to cause an elongation of the skull and opening of two sutures that run down the sides of the skull. Using measurements from a CT scan of the volume inside the skull at different areas and the distances between the bones in the cranial sutures, it is possible to diagnose sagittal synostosis. Data

Fig. 15.9 Types of suture in early human development.

set sagittal.dat|mat|xlsx contains measurements on 60 children, 20 with diagnosed sagittal synostosis and 40 normal controls (Walker, 2013; Credits to Dr. Barbara Boyan, Dr. Zvi Schwartz and Dr. Chris Hermann) The variables are described in the following table:

<table>
<thead>
<tr>
<th>Column</th>
<th>Variable</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>y</td>
<td>1 - Synostosis, 0 - Normal</td>
</tr>
<tr>
<td>2</td>
<td>x₁</td>
<td>Percent of volume in the front of the skull</td>
</tr>
<tr>
<td>3</td>
<td>x₂</td>
<td>Percent of sagittal suture that is open</td>
</tr>
<tr>
<td>4</td>
<td>x₃</td>
<td>Tangent distance in the left coronal suture</td>
</tr>
<tr>
<td>5</td>
<td>x₄</td>
<td>Tangent distance in the right coronal suture</td>
</tr>
</tbody>
</table>

(a) Model y by a logistic regression that uses x₁ and x₂ as predictors. Write down the model. Predict the probability of synostosis for a child with measured x₁ = 20 and x₂ = 61.
(b) If the predicted probability exceeds 0.5, decide \( \hat{y} = 1 \). What is the total number of errors made by your model, that is, how many predicted \( \hat{y}_i \)'s differ from the corresponding observed \( y_i \)?
(c) Show that the logistic model with all four variables \( x₁,…,x₄ \) as predictors makes no errors; that is, all \( \hat{y}_i \) are equal to \( y_i \).

15.19. **Shocks.** An experiment was conducted (Dalziel et al., 1941) to assess the effect of small electrical currents on farm animals, with the eventual goal of understanding the effects of high-voltage power lines on livestock. The
experiment was carried out with seven cows using six shock intensities, 0, 1, 2, 3, 4, and 5 milliamps (shocks on the order of 15 milliamps are painful for many humans). Each cow was given 30 shocks, 5 at each intensity, in random order. The entire experiment was then repeated, so each cow received a total of 60 shocks. For each shock, the response of mouth movement was either present or absent. The data as quoted give the total number of responses, out of 70 trials, at each shock level. We ignore cow differences and differences between blocks (experiments).

<table>
<thead>
<tr>
<th>Current (milliamps)</th>
<th>Number of responses</th>
<th>Number of trials</th>
<th>Proportion of responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>n</td>
<td>p</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>70</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>70</td>
<td>0.129</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>70</td>
<td>0.300</td>
</tr>
<tr>
<td>3</td>
<td>47</td>
<td>70</td>
<td>0.671</td>
</tr>
<tr>
<td>4</td>
<td>60</td>
<td>70</td>
<td>0.857</td>
</tr>
<tr>
<td>5</td>
<td>63</td>
<td>70</td>
<td>0.900</td>
</tr>
</tbody>
</table>

Using logistic regression and noninformative priors on its parameters, estimate the proportion of responses after a shock of 2.5 milliamps. Find 95% credible set for the population proportion.

15.20. **Separating Wild and Mutant** *C. elegans*. *C. elegans*, a soil dwelling nematode, is a highly studied multicellular organism that offers several experimental advantages such as a short life span, ease of culture, and transparency. Synapses in *C. elegans* can be imaged with fluorescent gene reporters such as GFP. San Miguel Delgadillo, et al. (2012) monitored presynaptic sites in motorneurons through a GFP-tagged protein located in synaptic vesicles. From these images, descriptors for the morphology of synaptic sites can be extracted to provide a phenotypic profile of the animal.

Researchers were faced with the problem of differentiating wild type and mutant *C. elegans* populations based on the very subtle phenotypic differences that can be presented at presynaptic sites, as quantified through 74 descriptors obtained through quantitative analysis of fluorescent images.

The data [celegans.dat][mat]xlsx (courtesy of Hang Lu Lab at Georgia Tech) contains 1,584 rows (observations) and 76 columns. The first 74 columns contain descriptors for synapse morphology (features), the 75th column is type of mutant (1–7, 0 is wild), and the last column contains an indicator of mutation (0 – wild, 1 – mutant).

Denote the matrix of features by $X$ and the indicator of mutation by $y$. The features $X_2, X_{20}, X_{37}, X_{51},$ and $X_{56}$ are selected as the most predictive columns in $X$ for determination of mutation.

(a) Fit the logistic regression
logit(Pr(y = 1)) = β₀ + β₁X₂ + β₂X₂₀ + β₃X₃₇ + β₄X₅₁ + β₅X₅₆ + ε.

(b) Using the fit from (a) estimate the probability that an observed C. elegans is a mutant if the recorded features were X₂ = 8.5, X₂₀ = 12.5, X₃₇ = 600, X₅₁ = 90, and X₅₆ = 0.5, respectively.

(c) Find the deviance and McFadden pseudo-R² for the model in (a).

(d) Describe in words (one paragraph) how you would classify C. elegans to mutant and wild using logistic regression. According to your description, how would you classify the C. elegans from (b)?

15.21. Ants. The data set `ants.csv` discussed in Gotelli and Ellison (2002), provides the ant species richness (number of ant species) found in 64-square-meter sampling grids in 22 forests (coded as 1) and 22 bogs (coded as 2) surrounding the forests in Connecticut, Massachusetts, and Vermont. The sites span 3° of latitude in New England. There are 44 observations on four variables (columns in data set): Ants – number of species, Habitat – forests (1) and bogs (2), Latitude, and Elevation – in meters above sea level.

(a) Using Poisson regression, model the number of ant species (Ants) with covariates Habitat and Elevation. Report the model coefficients and deviance.

(b) For a sampling grid unit located in a forest at the elevation of 100 m how many species the model from (a) predicts?

(c) Do the calculations from (a) and (b) using Open/WinBUGS with non-informative priors on all parameters. For the model coefficients and the prediction report 95% credible sets.

15.22. Sharp Dissection and Postoperative Adhesions Revisited. In Exercise 14.3 we fitted a linear relationship between the logarithm of the amount of sharp dissection laSD (predictor) and severity score sesco (response). Criticize this linear model. Model this relationship using Poisson regression and graphically compare the linear and Poisson fits.

15.23. Airfreight Breakage. A substance used in biological and medical research is shipped by air freight to users in cartons of 1,000 ampules. The data below, involving ten shipments, were collected on the number of times a carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y).

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>0</th>
<th>2</th>
<th>0</th>
<th>3</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>16</td>
<td>9</td>
<td>17</td>
<td>12</td>
<td>22</td>
<td>13</td>
<td>8</td>
<td>15</td>
<td>19</td>
<td>11</td>
</tr>
</tbody>
</table>

Using WinBUGS, fit Y by Poisson regression, with X as a covariate. According to your model, how many packages will be broken if the number of shipment routes is X = 4?
15.24. **Body Fat Affecting Accuracy of Heart Rate Monitors.** In the course *Problems in Biomedical Engineering I* at Georgia Tech, a team of students investigated whether the readings of heart rate from a chest strap monitor (Polar T31) were influenced by the subject's percentage of body fat. Hand counts facilitated by a stethoscope served as the gold standard. The absolute differences between device and hand counts (AD) were regressed on body fat (BF) measurements. The measurements for 28 subjects are provided below:

<table>
<thead>
<tr>
<th>Subj.</th>
<th>BF</th>
<th>AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17.8</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>13.2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>7.7</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>11.8</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>23.9</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>27.6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>27.6</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>18.8</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>13.4</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>39.4</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>6.8</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>12.5</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>19.9</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>19.9</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>25.1</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>18.3</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>16.9</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>6.8</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>36.0</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>31.9</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>17.4</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>24.1</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>12.9</td>
<td>2</td>
</tr>
<tr>
<td>24</td>
<td>30.1</td>
<td>6</td>
</tr>
<tr>
<td>25</td>
<td>17.1</td>
<td>4</td>
</tr>
<tr>
<td>26</td>
<td>18.4</td>
<td>4</td>
</tr>
<tr>
<td>27</td>
<td>14.6</td>
<td>4</td>
</tr>
<tr>
<td>28</td>
<td>26.8</td>
<td>3</td>
</tr>
</tbody>
</table>

A significant nonconstant representation of AD as a function of BF can be translated as a significant influence of percent body fat on the accuracy of the device.

(a) Why is linear regression not adequate here?
(b) Fit the Poisson regression $AD \sim \text{Po}(\exp(b_0 + b_1BF))$. Is the slope $b_1$ from the linear part significantly positive?
(c) Using WinBUGS find the 95% credible set for the slope $b_1$ in the linear part of a Bayesian Poisson regression model. Use non-informative priors.

15.25. **Micronuclei Assay.** The micronuclei (MN) assay procedure involves breaking the DNA of lymphocytes in a blood sample with a powerful dose of radiation, then measuring the efficiency of its ability to repair itself. Micronuclei are fragments of DNA that have not healed back into either of the two daughter nuclei after irradiation. The MN assay entails scoring the number of micronuclei; the higher the number, the less efficient is a person’s DNA repair system. The dose response of the number of micronuclei in cytokinesis-blocked lymphocytes after in-vitro irradiation of whole blood with X-rays in the dose range 0–4 Gy was studied by Thierens et al. (1991). The data provided in table are from one patient (male, 54 y.o.) and represent the frequency of micronuclei numbers for six levels of radiation, 0, 0.5, 1, 2, 3, and 4 Gy.
(a) Fit a Poisson regression in which the number of micronuclei is the response \( y \) and dose is a covariate \( x \). Plot Poisson intensity \( \lambda \) as a function of dose.

(b) What is the average number of micronuclei for dose of 3.5 Gy?

(c) Simulate 1,000 micronuclei counts for \( \lambda \) corresponding to dose 3.5 Gy. Summarize the simulation output.

**Hint:** Use MATLAB’s `glmfit` with ‘poisson’ distribution and ‘log’ link. This will estimate the regression coefficients needed for (b–c). For plotting, use `glmval`.

The data need to be appropriately recoded and one way to do so is shown below:

```matlab
y=[ zeros(976,1); ones(21,1); 2*ones(3,1); ... zeros(936,1); ones(61,1); 2*ones(3,1); ... zeros(895,1); ones(94,1); 2*ones(11,1); ... zeros(760,1); ones(207,1); 2*ones(32,1); 3*ones(1,1); ... zeros(583,1); ones(302,1); 2*ones(97,1); 3*ones(12,1); 4*ones(6,1); ... zeros(485,1); ones(319,1); 2*ones(147,1); 3*ones(35,1); 4*ones(11,1); ... 5*ones(2,1); 6*ones(1,1)];
```

```matlab
x = [zeros(976+21+3,1); ... 0.5 * ones(936+61+3,1); ... ones(895+94+11,1); ... 2*ones(760+207+32+1,1); ... 3*ones(583+302+97+12+6,1); ... 4*ones(485+319+147+35+11+2+1,1)];
```

15.26. **Miller Lumber Company Customer Survey.** Kutner et al. (2005) analyze a data set from a survey of customers of the Miller Lumber Company. The response is the total number of customers (in a representative 2-week period) coming from a tract of a metropolitan area within 10 miles from the store. The covariates include five variables concerning the tracts: number of housing units, average income in dollars, average housing unit age in years, distance to nearest competitor in miles, and distance to store in miles. Fit and assess a Poisson regression model for the number of customers as predicted by the covariates. The data are in `lumber.m`.

15.27. **SO₂, NO₂, and Hospital Admissions.** Fan and Chen (1999) discuss a public health data set consisting of daily measurements of pollutants
and other environmental factors in Hong Kong between January 1, 1994 and December 31, 1995. The association between levels of pollutants and the number of daily hospital admissions for circulation and respiratory problems is of particular interest.

The data file hospitaladmissions.dat consists of six columns: (1) year, (2) month, (3) day in month, (4) concentration of sulfur dioxide SO\textsubscript{2}, (5) concentration of pollutant nitrogen NO\textsubscript{2}, and (6) daily number of hospital admissions.

(a) Using logistic regression, determine how the probability of a high level of sulfur dioxide (with values > 20 \(\mu\text{g/m}^3\)) is associated with the level of pollutant nitrogen NO\textsubscript{2}.

(b) Using a Poisson regression model, explore how the expected number of hospital admissions varies with the level of NO\textsubscript{2}.

(c) Suppose that on a particular day the level of NO\textsubscript{2} was measured at 100. Estimate the probability of a high level of sulfur dioxide in (a) and the expected number of hospital admissions in (b).

15.28. Kidney Stones. Charig et al. (1986) provide data on the success rates of two methods of treating kidney stones: open surgery methods and percutaneous nephrolithotomy. There are two predictors: size of stone and method. Size of stone is set at two levels: < 2 cm in diameter, coded as Small, and > 2 cm in diameter, coded as Large. The two methods are coded as A (open surgery) and B (percutaneous nephrolithotomy). The outcome of interest is the outcome of the treatment (Success, Failure).

<table>
<thead>
<tr>
<th>Count</th>
<th>Size</th>
<th>Method</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>Small</td>
<td>A</td>
<td>Success</td>
</tr>
<tr>
<td>6</td>
<td>Small</td>
<td>A</td>
<td>Failure</td>
</tr>
<tr>
<td>234</td>
<td>Small</td>
<td>B</td>
<td>Success</td>
</tr>
<tr>
<td>36</td>
<td>Small</td>
<td>B</td>
<td>Failure</td>
</tr>
<tr>
<td>192</td>
<td>Large</td>
<td>A</td>
<td>Success</td>
</tr>
<tr>
<td>71</td>
<td>Large</td>
<td>A</td>
<td>Failure</td>
</tr>
<tr>
<td>55</td>
<td>Large</td>
<td>B</td>
<td>Success</td>
</tr>
<tr>
<td>25</td>
<td>Large</td>
<td>B</td>
<td>Failure</td>
</tr>
</tbody>
</table>

There are four combinations of the covariates: Small A, Small B, Large A, and Large B. Find the relative frequencies of the outcome “Success” and compare them with the model-predicted probabilities using logistic regression.

Show that these data hide Simpson’s paradox.
MATLAB AND WINBUGS FILES AND DATA SETS USED IN THIS CHAPTER

http://statbook.gatech.edu/Ch15.Logistic/

arrhythmia.m, arthritis2.m, beetleBliss1.m, beetleBliss2.m, bumpus.m,
ciaesarean.m, cellular.m, counterr.m, dmdreg.m, ihga.m, kyphosis.m,
logisticmle.m, logisticSeed.m, lumber.m, outbreak.m, ponv.m,
prostatecanc.m, sagittal.m, UptonFingleton.m

accidentssimple.odc, arrhythmia.odc, beetles.odc, bliss.odc,
bystanders.odc, caesarean.odc, celldifferentiation.odc, errors1.odc,
geriatric.odc, microdamage.odc, raynaud.odc, remission.odc,
tribolium.odc, tromboembolism.odc

ants.dat, arrhythmia.mat|xlsx, arrhythmiadata.m, arthritis2.dat,
birthweight.dat, bumpus.mat, cardiac.mat|txt, celegans.dat|mat|xlsx,
ceular.dat, diabetes.dat, dmd.dat|mat, finney.dat,
hospitaladmissions.dat, ihga.dat, kyphosis.dat|txt, lowbwt.dat,
microdamage.dat|mat, outbreak.mat, pima.dat, PONV.mat|xls, programm.mat,
prostatecanc.dat, sagittal.mat|xlsx, tribolium.mat

CHAPTER REFERENCES


Chapter 16
Inference for Censored Data and Survival Analysis

The first condition of progress is the removal of censorship.
– George Bernard Shaw

WHAT IS COVERED IN THIS CHAPTER

- Parametric Models for Time-to-Event Data
- Kaplan-Meier Estimator, Mantel’s Logrank Test
- Cox Proportional Hazards Model
- Bayesian Approaches

16.1 Introduction

Survival analysis models the survival times of a group of subjects, usually with some kind of medical condition, and generates a survival curve, which shows how many of the subjects are “alive” or survive over time.

What makes survival analysis different from standard regression methodology is the presence of censored observations; in addition, some subjects may leave the study and will be lost to follow-up. Such subjects were known to have survived for some amount of time (up until the time we last saw them), but we do not know how much longer they might ultimately have survived. Several methods have been developed for using this “at least this
long” information to finding unbiased survival curve estimates, the most popular being the nonparametric method of Kaplan and Meier.

An observation is said to be censored if we know only that it is less than (or greater than) a certain known value. For instance, in clinical trials, one could be interested in patients’ survival times. Survival time is defined as the length of time between diagnosis and death, although other “start” events, such as surgery, and other “end” events, such as relapse of disease, increase in tumor size beyond a particular threshold, rejection of a transplant, etc., are commonly used. Because of many constraints, trials cannot be run until the endpoints are observed for all patients. Just because for a particular subject the time to endpoint is not fully observed, partial information is still available: the patient survived up to the end of the observational period, and this should be incorporated into the analysis. Such observations are called right censored. Observations can also be left censored, for example, an assay may have a detection threshold. In order to utilize information contained in censored observations, special methods of analysis are required.

Engineers are often interested in the reliability of various devices, and most of the methodology from survival analysis is applicable in device reliability analyses. There are of course important differences. In the reliability of multicomponent systems an important aspect is optimization of the number and position of components. Analogous considerations with organs or parts of organs as components in living systems (animals or humans) are impossible. Methods such as “accelerated life testing” commonly used in engineering reliability are inappropriate when dealing with human subjects.

However, in comparing the lifetimes of subjects in clinical trials involving different treatments (humans, animals) or different engineering interventions (systems, medical devices), the methodology that deals with censored observations is shared.

### 16.2 Definitions

Let $T$ be a continuous random variable with CDF $F(t)$ representing a lifetime. The survival (or survivor) function is the tail probability for $T$, expressed as $S(t) = 1 - F(t)$, $t > 0$. The function $S(t)$ gives the probability of surviving up to time $t$, that is,

\[
S(t) = \mathbb{P}(T > t).
\]
The hazard function or hazard rate is defined as

\[ h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}, \]

when \( T \) has a density \( f(t) \). Note that \( S'(t) = -f(t) \). It is insightful to represent \( h(t) \) in limit terms,

\[
\frac{S(t) - S(t + \Delta t)}{\Delta t} \times \frac{1}{S(t)} = \frac{F(t + \Delta t) - F(t)}{\Delta t} \times \frac{1}{S(t)} \\
= \frac{\mathbb{P}(t < T \leq t + \Delta t)}{\Delta t} \frac{\mathbb{P}(T > t)}{\mathbb{P}(T > t)} \\
= \frac{\mathbb{P}(T \leq t + \Delta t | T > t)}{\Delta t}
\]

when \( \Delta t \to 0 \). It represents an instantaneous probability that an event that was not observed up to time \( t \) will be observed before \( t + \Delta t \), when \( \Delta t \to 0 \).

Cumulative hazard is defined as

\[ H(t) = \int_0^t h(s)ds. \]

Both hazard and cumulative hazard uniquely determine the distribution of lifetime \( F \),

\[ F(t) = 1 - \exp \left\{ - \int_0^t h(s)ds \right\} = 1 - \exp \{-H(t)\}. \]

Cumulative hazard can also be connected to the survival function as \( H(t) = -\log S(t) \), or

\[ S(t) = \exp \{-H(t)\}. \quad (16.1) \]

**Example 16.1. Constant Hazard.** The hazard function for an exponential distribution with density \( f(t) = \lambda e^{-\lambda t}, t \geq 0, \lambda > 0 \) is constant in time, \( h(t) = \lambda \).

**Example 16.2. Linear Hazard.** Identify distributions for which the hazard rate is linear, \( h(t) = a + bt, t \geq 0 \).

Since

\[ H(t) = \int_0^t h(u)du = at + \frac{bt^2}{2}, \]
806 16 Inference for Censored Data and Survival Analysis

according to (16.1),

\[
S(t) = 1 - F(t) = \exp \left\{ -at - \frac{bt^2}{2} \right\},
\]

and

\[
f(t) = (a + bt) \exp \left\{ -at - \frac{bt^2}{2} \right\}.
\]

This represents a density if \( b \geq 0 \). For \( a = 0 \) the above is the Rayleigh distribution, and for \( b = 0 \) it is the exponential.

Example 16.3. Weibull’s Hazard. The hazard rate for a one-parameter Weibull distribution with CDF \( F(t) = 1 - \exp\{-t^\gamma\} \) and density \( f(t) = \gamma t^{\gamma-1} \exp\{-t^\gamma\}, \ t \geq 0, \gamma > 0 \) is \( h(t) = \gamma t^{\gamma-1} \). The parameter \( \gamma \) is called a shape parameter. Depending on the shape parameter \( \gamma \), the hazard function \( h(t) \) could model various types of survival analyses.

The two-parameter version of Weibull that was introduced in Section 5.5.9 is used more frequently. It is defined as \( F(t) = 1 - \exp\{-\lambda t^\gamma\} \), with density \( f(t) = \lambda \gamma t^{\gamma-1} \exp\{-\lambda t^\gamma\}, \ t \geq 0, \gamma > 0, \lambda > 0 \). The parameter \( \lambda \) is called a rate parameter. In this case \( h(t) = \lambda \gamma t^{\gamma-1} \). Figure 16.1 shows hazard functions for different values of shape parameters \( \gamma \). For \( \gamma = 1 \), the hazard is constant (Weibull distribution becomes exponential), for \( \gamma < 1 \), the hazard is decreasing, and for \( \gamma > 1 \), it is increasing.

![Fig. 16.1 Hazard function of two parameter Weibull distribution with \( \lambda = 1 \) and \( \gamma = 0.9, 1, \) and 1.1.](image-url)
Two important summaries in the parametric case (where the survival distribution is specified up to a parameter) are mean residual life (mrl) and median life, defined, respectively, as

\[
mrl(t) = \frac{\int_t^\infty S(x)dx}{S(t)},
\]
\[
t_{0.5} : S(t_{0.5}) = 0.5.
\]

The hazard rate \( h(t) \) and mean residual life \( mrl(t) \) are connected,

\[
h(t) = \frac{1 + (mrl(t))'}{mrl(t)}.
\]

Example 16.4. Exponential Mean Residual Life. For an exponential lifetime, the expected lifetime and mrl coincide. Indeed, \( \mathbb{E}T = 1/\lambda \), while

\[
mrl(t) = \frac{\int_t^\infty e^{-\lambda x}dx}{e^{-\lambda t}} = \frac{1}{\lambda}.
\]

At first glance this looks like a paradox; the expected total lifetime \( \mathbb{E}T \) is equal to the residual lifetime \( mrt(t) \) regardless of \( t \). This is an example of the “inspection paradox” that follows from the memoryless property of the exponential distribution, page 201.

The median lifetime of an exponential distribution is \( t_{0.5} = (\log 2)/\lambda \).

Example 16.5. Estimation Parameters in Weibull Distribution by Regression. Recall that the CDF of a two-parameter Weibull distribution is

\[
F(x) = 1 - e^{-\lambda x^r},
\]

where \( \lambda \) is rate parameter and \( r \) is the shape parameter. Given a sample \( X_1, X_2, \ldots, X_n \), the goal is to estimate \( \lambda \) and \( r \).

The MLE equations do not have solutions in closed form and numerical approximations are used. Simpler and often superior estimation is based on a combination of method of moments and linear regression.

Note that
\[ 1 - F(x) = e^{-\lambda x^r} \]
\[ \log(1 - F(x)) = -\lambda x^r \]
\[ \log \left( \frac{1}{1 - F(x)} \right) = \lambda x^r. \]

After taking the logarithm, the equation becomes,
\[ \log \left( \log \left( \frac{1}{1 - F(x)} \right) \right) = \log(\lambda) + r \log x. \]

If \( X(i) \) is \( i \)th order statistic from \( X_1, \ldots, X_n \) coming from distribution with a CDF \( F \), it holds that
\[ \mathbb{E}F(X(i)) = \frac{i}{n+1}. \]
This is true for any distribution, since \( F(X(i)) = U(i) \) where \( U(i) \) is the \( i \)th order statistic from uniform \( U(0,1) \) distribution, and \( \mathbb{E}U(i) = i/(n+1) \).

The idea is to calculate regression responses \( y_i \) using \( \mathbb{E}F(X(i)) \) in place of \( F(X(i)) \),
\[ y_i = \log \left( \log \left( \frac{1}{1 - \mathbb{E}F(X(i))} \right) \right) = \log \left( \log \left( \frac{n + 1}{n + 1 - i} \right) \right), \]
and regress the responses against ordered observations \( x_i = X(i) \).

If \( b_0 \) and \( b_1 \) are estimated intercept and slope in this regression, then
\[ \hat{\lambda} = e^{b_0} \quad \text{and} \quad \hat{r} = b_1. \]

The following MATLAB script simulates 2,000 observations from Weibull \( \text{Wei}(2,3) \) distribution. The estimation of parameters \( r \) and \( \lambda \) is done by regression, and estimators are compared results from built in MATLAB’s \texttt{weibfit}.

```matlab
% estimweibull.m
s = RandStream('mt19937ar','Seed',0);
RandStream.setGlobalStream(s);

lambda=3;
r=2;
% Simulate 2000 Wei(r=2, lambda=3)
n = 2000;
xx = wblrnd(lambda^(-1/r), r, [n 1]);
% In MATLAB’s parametrization of Weibull distribution
% the scale parameter is lambda^(-1/r)
x = sort(xx); i=(1:n)';
y = log(log((n+1)/(n+1 - i)));
```

\[ 808 \]
16.3 Inference with Censored Observations

We will consider the case of right-censored data (the most common type of censoring) and two approaches: a parametric approach, in which the survival function $S(t)$ would have a specific functional form, and a non-parametric approach, in which no such functional form is assumed.

### 16.3.1 Parametric Approach

In the parametric approach the models depend on the parameters, and the parameters are estimated by taking into account both uncensored and censored observations. We will show how to find an MLE in the general case and illustrate it on an exponential lifetime distribution.

Let $(t_i, \delta_i), i = 1, \ldots, n$ be observations of a lifetime $T$ for $n$ individuals, with $\delta \in \{0, 1\}$ indicating fully observed and censored lifetimes, and let $k$ observations be fully observed while $n - k$ are censored. Suppose that the underlying lifetime $T$ has a density $f(t|\theta)$ with survival function $S(t|\theta)$. Then the likelihood is

$$ L(\theta|t_1, \ldots, t_n) = \prod_{i=1}^{n} \left( h(t_i|\theta)^{1-\delta_i} \times (S(t_i|\theta))^{\delta_i} \right) = \prod_{i=1}^{k} f(t_i|\theta) \times \prod_{i=k+1}^{n} S(t_i|\theta). $$

Since $h(t_i|\theta) \times (S(t_i|\theta)) = f(t_i|\theta)$, then

$$ L(\theta|t_1, \ldots, t_n) = \prod_{i=1}^{n} \left( h(t_i|\theta)^{1-\delta_i} \times S(t_i|\theta) \right). \quad (16.2) $$

**Example 16.6. MLE for Censored Exponential Lifetimes.** We will show that for an exponential lifetime in the presence of right-censoring, the MLE for
\[ \hat{\lambda} = \frac{k}{\sum_{i=1}^{n} t_i}, \]  

(16.3)

where \( k \) is the number of noncensored data points and \( \sum_{i=1}^{n} t_i \) is the sum of all observed and censored times. From (16.2), the likelihood is \( L = \lambda^k \exp\{-\lambda \sum_{i=1}^{n} t_i\} \). By taking the log and differentiating, we get the MLE as the solution to \( \frac{k}{\lambda} - \sum_{i=1}^{n} t_i = 0 \).

The variance of the MLE \( \hat{\lambda} \) is \( \frac{k}{(\sum_{i=1}^{n} t_i)^2} \) and can be used to find the confidence interval for \( \lambda \) (Exercise 16.2).

**Example 16.7. Immunoperoxidase and BC.** Data analyzed in Sedmak et al. (1989) and also in Klein and Moeschberger (2003) represent times to death (in months) for breast cancer patients with different immunohistochemical responses. Out of 45 patients in the study, 9 were immunoperoxidase positive while the remaining 36 were negative (+ denotes censored time).

<table>
<thead>
<tr>
<th>Immunoperoxidase negative</th>
<th>Immunoperoxidase positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>19, 25, 30, 34, 37, 46, 47, 51, 56, 57, 61, 66, 67, 74, 78, 86, 122+,...</td>
<td>22, 23, 38, 42, 73, 77, 89, 115, 144+</td>
</tr>
</tbody>
</table>

Assume that lifetimes are exponentially distributed and that rates \( \lambda_1 \) (for Immunoperoxidase negative) and \( \lambda_2 \) (for Immunoperoxidase positive) are to be estimated. The following MATLAB code finds MLEs of \( \lambda_1 \) and \( \lambda_2 \), first directly by using (16.3) and then by using MATLAB’s built-in function `mle` with option ‘censoring’. 

```matlab
```
16.3 Inference with Censored Observations

\[ \text{ImmPeroxNeg} = \{19, 25, 30, 34, 37, 46, 51, 56, 57, 61, 66, 67, 74, 78, 86, \ldots \} \]
\[ \text{CensorIPN} = \{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1 \} \]
\[ \text{ImmPeroxPos} = \{22, 23, 38, 42, 73, 77, 89, 115, 144\} \]
\[ \text{CensorIPP} = \{0, 0, 0, 0, 1\} \]

\[ k_1 = \text{sum}(1-\text{CensorIPN}) \] 16

\[ k_2 = \text{sum}(1-\text{CensorIPP}) \] 8

\[ \text{MLEs of rate } \lambda \text{ for 2 samples.} \]
\[ \hat{\lambda}_1 = \frac{k_1}{\text{sum}(\text{ImmPeroxNeg})} \] 0.0042
\[ \hat{\lambda}_2 = \frac{k_2}{\text{sum}(\text{ImmPeroxPos})} \] 0.0128

\[ [\text{reclambdahat1} \ \text{lambc1}] = \text{mle}((\text{ImmPeroxNeg}), \ldots) \]
\[ \text{distribution} = \text{exponential}, \text{censoring} = \text{CensorIPN} \]
\[ \% 237.6250 \]
\[ \% 153.6769 \]
\[ \text{reclambdahat2} \] 77.8750
\[ \% 43.1959 \]

\[ \% (\text{MATLAB parametrization}) \] scale to rate
\[ \lambda_1 = \frac{1}{\text{reclambdahat1}} \] 0.0042
\[ \lambda_2 = \frac{1}{\text{reclambdahat2}} \] 0.0128

As is indicated by the code, the patients who are immunoperoxidase positive are at increased risk since the rate \( \hat{\lambda}_2 = 0.0128 \) exceeds \( \hat{\lambda}_1 = 0.0042 \).

In MATLAB, \texttt{dfittool}, \texttt{normfit}, \texttt{wblfit}, and other commands for fitting parametric distributions can be applied to censored data by specifying the censoring vector at the input.

### 16.3.2 Nonparametric Approach: Kaplan–Meier or Product–Limit Estimator

Assume that individuals in the study are assessed at discrete time instances \( t_1, t_2, \ldots, t_k \), which may not be equally spaced. Typically, the times \( t_i \) are selected when failures occur. If we want to calculate the probability of survival up to time \( t_i \), then by the chain rule of conditional probabilities and their Markovian property,
\[ \hat{S}(t_i) = \mathbb{P}( \text{surviving to time } t_i ) = \mathbb{P}( \text{survived up to time } t_1 ) \]
\[ \times \mathbb{P}( \text{surviving to time } t_2 \mid \text{survived up to time } t_1 ) \]
\[ \times \mathbb{P}( \text{surviving to time } t_3 \mid \text{survived up to time } t_2 ) \]
\[ \times \mathbb{P}( \text{surviving to time } t_i \mid \text{survived up to time } t_{i-1} ). \]

It is assumed that \( t_0 = 0 \).

Suppose that \( r_i \) subjects are at risk at time \( t_{i-1} \) and are not censored at time \( t_{i-1} \). In the \( i \)th interval \([t_{i-1}, t_i)\) among these \( r_i \) subjects \( d_i \) have an event, \( \ell_i \) are censored, and \( r_{i+1} \) survive. The \( r_{i+1} \) subjects will be at risk at the beginning of the \((i+1)\)th time interval \([t_i, t_{i+1})\), that is, at time \( t_i \). Thus, \( r_i = d_i + \ell_i + r_{i+1} \). We can estimate the probability of survival up to time \( t_i \), given that one survived up to time \( t_{i-1} \), as \( 1 - d_i / (r_{i+1} + d_i + \ell_i) = 1 - d_i / r_i \).

The \( \ell_i \) subjects censored at time \( t_i \) do not contribute to the survival function for times \( t > t_i \).

\[ \hat{S}(t) = \left( 1 - \frac{d_1}{r_1} \right) \times \left( 1 - \frac{d_2}{r_2} \right) \times \cdots \times \left( 1 - \frac{d_i}{r_i} \right) \]
\[ = \prod_{t_i \leq t} \left( 1 - \frac{d_i}{r_i} \right), \text{ for } t \geq t_1; \]
\[ \hat{S}(t) = 1, \text{ for } t < t_1. \]

This is the celebrated Kaplan–Meier or product-limit estimator (Kaplan and Meier, 1958). This result has been one of the most influential developments in the past century in statistics; the paper by Kaplan and Meier is the most cited paper in the field of statistics (Stigler, 1994).

For uncensored observations, the Kaplan–Meier estimator is identical to the complement of the empirical CDF. The difference occurs when there is a censored observation – then the Kaplan–Meier estimator takes the “weight” normally assigned to that observation and distributes it evenly among all observed values to the right of the censored observation. This is intuitive because we know that the true value of the censored observation must be somewhere to the right of the censored value, but information about what the exact value should be is lacking. Thus all observed values larger than the censored observation are treated in the same way.

The variance of Kaplan–Meier estimator is estimated by Greenwood’s formula (Greenwood, 1926):

\[ \tau^2_S(t) = (\hat{S}(t))^2 \times \sum_{t_i \leq t} \frac{d_i}{r_i(r_i - d_i)}. \]
16.3 Inference with Censored Observations

The pointwise confidence intervals (for a fixed time \( t^* \)) for the survival function \( S(t^*) \) can be found in several ways. The most popular confidence intervals are

**linear**

\[
\left[ \hat{S}(t^*) - z_{1-\alpha/2} \tau S(t^*), \hat{S}(t^*) + z_{1-\alpha/2} \tau S(t^*) \right],
\]

**log-transformed**

\[
\left[ (\hat{S}(t^*)) \exp\left\{ \frac{-z_{1-\alpha/2} \tau S(t^*)}{S(t^*)} \right\}, (\hat{S}(t^*)) \exp\left\{ \frac{z_{1-\alpha/2} \tau S(t^*)}{S(t^*)} \right\} \right],
\]

and **log-log-transformed**

\[
\left[ (\hat{S}(t^*))^v, (\hat{S}(t^*))^{1/v} \right], \quad v = \exp\left\{ \frac{z_{1-\alpha/2} \tau S(t^*)}{S(t^*) |\log \hat{S}(t^*)|} \right\}.
\]

Although not centered at \( \hat{S}(t^*) \), the log- and log-log-transformed intervals are considered superior to the linear. This is because \( \hat{S}(t^*) \) is not well approximated by a normal distribution, especially when \( S(t^*) \) is close to 0 or 1.

The pointwise confidence intervals given above differ from simultaneous confidence bounds on \( S(t) \) for which the confidence of \( 1 - \alpha \) means that the probability that *any* part of the curve \( S(t) \) will fall outside the bounds does not exceed \( \alpha \). Such general bounds are naturally wider than those generated by pointwise confidence intervals, since the overall confidence is controlled. Two important types of confidence bands are Nair's equal precision bands and the Hall–Wellner bands. Description of these bounds are beyond the scope of this text; see Klein and Moeschberger (2003, p. 109), for further discussion and implementation. The bounds computed in MATLAB’s \([f,t,fl,fu]=ecdf(...)\) also return lower and upper confidence bounds for the CDF. These bounds are calculated using Greenwood’s formula and are not simultaneous confidence bounds.

The Kaplan–Meier estimator also provides an estimator for the cumulative hazard \( H(t) \) as

\[
\hat{H}(t) = -\log (\hat{S}(t)).
\]

Better small-sample performance in estimating the cumulative hazard can be achieved by the Nelson–Aalen estimator,
\[ H(t) = \begin{cases} 0, & \text{for } t \leq t_1 \\ \sum_{i=1}^{t} d_i / r_i, & \text{for } t > t_1 \end{cases} \]

with an estimated variance \( \sigma^2_H(t) = \sum_{i=1}^{t} d_i / r_i^2 \). By using \( \tilde{H}(t) \) and \( \sigma^2_H(t) \), pointwise confidence intervals on \( H(t) \) can be obtained.

**Example 16.8.** Catheter Complications in Peritoneal Dialysis. The following example is from Chadha et al. (2000). The authors studied a sample of 36 pediatric patients undergoing acute peritoneal dialysis through Cook catheters. They wished to examine how long these catheters performed properly. They noted the date of complication (either occlusion, leakage, exit-site infection, or peritonitis).

Half of the subjects had no complications before the catheter was removed. Reasons for removal of the catheter in this group of patients were that the patient recovered \( (n = 4) \), the patient died \( (n = 9) \), or the catheter was changed to a different type electively \( (n = 5) \). If the catheter was removed prior to complications, that represented a censored observation, because they knew that the catheter remained complication free at least until the time of removal.

<table>
<thead>
<tr>
<th>Day</th>
<th>At Risk, ( r_i )</th>
<th>Censored, ( \ell_i )</th>
<th>Fail, ( d_i )</th>
<th>( 1 - \frac{d_i}{r_i} )</th>
<th>KM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1 - 2/36 = 0.944</td>
<td>0.9444</td>
</tr>
<tr>
<td>2</td>
<td>36 - 8 - 2 = 26</td>
<td>2</td>
<td>2</td>
<td>1 - 2/26 = 0.92</td>
<td>0.92 \cdot 0.944 = 0.8718</td>
</tr>
<tr>
<td>3</td>
<td>26 - 2 - 2 = 22</td>
<td>1</td>
<td>2</td>
<td>1 - 2/22 = 0.91</td>
<td>0.91 \cdot 0.872 = 0.7925</td>
</tr>
<tr>
<td>4</td>
<td>22 - 1 - 2 = 19</td>
<td>1</td>
<td>1</td>
<td>1 - 1/19 = 0.95</td>
<td>0.95 \cdot 0.793 = 0.7508</td>
</tr>
<tr>
<td>5</td>
<td>19 - 1 - 1 = 17</td>
<td>6</td>
<td>3</td>
<td>1 - 3/17 = 0.82</td>
<td>0.6183</td>
</tr>
<tr>
<td>6</td>
<td>17 - 6 - 3 = 8</td>
<td>0</td>
<td>2</td>
<td>1 - 2/8 = 0.75</td>
<td>0.4637</td>
</tr>
<tr>
<td>7</td>
<td>8 - 0 - 2 = 6</td>
<td>0</td>
<td>1</td>
<td>1 - 1/6 = 0.83</td>
<td>0.3865</td>
</tr>
<tr>
<td>10</td>
<td>6 - 0 - 1 = 5</td>
<td>0</td>
<td>2</td>
<td>1 - 2/5 = 0.60</td>
<td>0.2319</td>
</tr>
<tr>
<td>12</td>
<td>5 - 0 - 2 = 3</td>
<td>0</td>
<td>1</td>
<td>1 - 2/3 = 0.33</td>
<td>0.0773</td>
</tr>
<tr>
<td>13</td>
<td>3 - 0 - 2 = 1</td>
<td>0</td>
<td>1</td>
<td>1 - 1/1 = 0.00</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

MATLAB script `chada.m` finds the Kaplan–Meier estimator and generates Figure 16.2. plots

```matlab
%chada.m

times=[1,1,1,1,1,1,1,1,1,2,2,2,...
   3,3,3,4,4,5,5,5,5,5,5,5,...
   6,6,7,10,10,12,12,13];
censored = [1,1,1,1,1,1,1,0,0,1,1,...
   0,0,1,0,0,1,0,1,1,1,1,...
   1,0,0,0,0,0,0,0,0,0,0];

% Calculate and plot KM estimator
ple(times, censored)
```
Example 16.9. **Strength of Weathered Cord.** Data from Crowder et al. (1991) lists strength measurements (in coded units) for 48 pieces of weathered cord. Seven of the pieces of cord were damaged and yielded strength measurements that are considered right-censored. That is, because the damaged cord was taken off the test, we know only the lower limit of its strength. In the MATLAB code below, the vector `data` represents the strength measurements, and the vector `censor` indicates (with a zero) if the corresponding observation in `data` is censored.

```matlab
data = [36.3, 41.7, 43.9, 49.9, 50.1, 50.8, 51.9, 52.1, 52.3, 52.3, ...
       52.4, 52.6, 52.7, 53.1, 53.6, 53.6, 53.9, 53.9, 54.1, 54.6, ...
       54.8, 54.8, 55.1, 55.4, 55.9, 56.0, 56.1, 56.5, 56.9, 57.1, ...
       57.1, 57.3, 57.7, 57.8, 58.1, 58.9, 59.0, 59.1, 59.6, 60.4, ...
       60.7, 62.8, 62.9, 63.4, 63.5, 64.0, 64.1, 64.2];
censor = [zeros(1,41), ones(1,7)];
[table] = ple(data, censor);
```

The table below shows how the Kaplan–Meier estimator is calculated for the first 16 measurements, which includes 7 censored observations. Figure 16.3 shows the estimated survival function for the cord strength data.
Fig. 16.3 Kaplan–Meier estimator cord strength (in coded units).

<table>
<thead>
<tr>
<th>$t_i$</th>
<th>$r_i$</th>
<th>$d_i$</th>
<th>$\ell_i$</th>
<th>$1 - \frac{d_i}{r_i}$</th>
<th>$\hat{S}(t_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>48</td>
<td>0</td>
<td>4</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>36.3</td>
<td>44</td>
<td>1</td>
<td>1</td>
<td>0.9773</td>
<td>0.9773</td>
</tr>
<tr>
<td>41.7</td>
<td>42</td>
<td>1</td>
<td>2</td>
<td>0.9762</td>
<td>0.9540</td>
</tr>
<tr>
<td>43.9</td>
<td>39</td>
<td>1</td>
<td>0</td>
<td>0.9744</td>
<td>0.9295</td>
</tr>
<tr>
<td>49.9</td>
<td>38</td>
<td>1</td>
<td>0</td>
<td>0.9737</td>
<td>0.9051</td>
</tr>
<tr>
<td>50.1</td>
<td>37</td>
<td>1</td>
<td>0</td>
<td>0.9730</td>
<td>0.8806</td>
</tr>
<tr>
<td>50.8</td>
<td>36</td>
<td>1</td>
<td>0</td>
<td>0.9722</td>
<td>0.8562</td>
</tr>
<tr>
<td>51.9</td>
<td>35</td>
<td>1</td>
<td>0</td>
<td>0.9714</td>
<td>0.8317</td>
</tr>
<tr>
<td>52.1</td>
<td>34</td>
<td>1</td>
<td>0</td>
<td>0.9706</td>
<td>0.8072</td>
</tr>
<tr>
<td>52.3</td>
<td>33</td>
<td>2</td>
<td>0</td>
<td>0.9394</td>
<td>0.7583</td>
</tr>
<tr>
<td>52.4</td>
<td>31</td>
<td>1</td>
<td>0</td>
<td>0.9677</td>
<td>0.7338</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>57.8</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0.8750</td>
<td>0.1712</td>
</tr>
<tr>
<td>58.1</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0.8571</td>
<td>0.1468</td>
</tr>
<tr>
<td>58.9</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0.8333</td>
<td>0.1223</td>
</tr>
<tr>
<td>59.0</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0.8000</td>
<td>0.0978</td>
</tr>
<tr>
<td>59.1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0.7500</td>
<td>0.0734</td>
</tr>
<tr>
<td>59.6</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0.6667</td>
<td>0.0489</td>
</tr>
<tr>
<td>60.4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0.5000</td>
<td>0.0245</td>
</tr>
<tr>
<td>60.7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
16.3.3 Comparing Survival Curves

In clinical trials it is often important to compare survival curves calculated for cohorts undergoing different treatments. Often one is interested in comparing the new treatment to the existing one or to a placebo. In comparing two survival curves, we are testing whether the corresponding hazard functions \( h_1(t) \) and \( h_2(t) \) coincide:

\[
H_0 : h_1(t) = h_2(t) \quad \text{versus} \quad H_1 : h_1(t) >, \neq, < h_2(t).
\]

The simplest comparison involves exponential lifetime distributions where the comparison between survival/hazard functions is simply a comparison of constant rate parameters. The statistic is calculated using logarithms of hazard rates as

\[
Z = \log \lambda_1 - \log \lambda_2 \sqrt{1/k_1 + 1/k_2},
\]

where \( k_1 \) and \( k_2 \) are numbers of observed (uncensored) survival times in the two comparison groups. Now, the inference relies on the fact that statistic \( Z \) is approximately standard normal.

Example 16.10. Comparing the Rates. In Example 16.7 the rate for immunoperoxidase-positive patients was larger than that of immunoperoxidase-negative patients. Was this difference significant?

\[
z = (\log(\hat{\lambda}_1) - \log(\hat{\lambda}_2))/\sqrt{1/k_1 + 1/k_2} \quad \% -2.5763
\]

\[
p = \text{normcdf}(z) \quad \% 0.0050
\]

As is evident from the code, the hazard rate \( \lambda_2 \) (and, in the case of exponential distribution, hazard function) for the immunoperoxidase-positive patients is significantly larger than the rate for negative patients, \( \lambda_1 \), with a \( p \)-value of half a percent.

Logrank Test. The logrank test compares survival functions in a nonparametric fashion. It was proposed by Mantel (1966) and applies Haenszel-Mantel theory on survival data in the form of \( 2 \times 2 \) tables.

Let \((r_{11}, d_{11}), (r_{12}, d_{12}), \ldots, (r_{1k}, d_{1k})\) be the number of people at risk and the number of people who died at times \( t_{11}, t_{12}, \ldots, t_{1k} \) in the first cohort, and \((r_{21}, d_{21}), (r_{22}, d_{22}), \ldots, (r_{2m}, d_{2m})\) be the number of people at risk and the number of people who died at times \( t_{21}, t_{22}, \ldots, t_{1m} \) in the second cohort. We merge the two data sets together with the corresponding times. Thus, there will be \( D = k + m \) time points if there are no ties, and each time point corresponds to a death from either the first or second cohort. For example, if times of events in the first sample are 1, 4, and 10 and in the second 2, 3, 7, and 8, then in the merged data sets the times will be 1, 2, 3, 4, 7, 8, and 10.
For a time \( t_i \) from the merged data set, let \( r_{1i} \) and \( r_{2i} \) correspond to the number of subjects at risk in cohorts 1 and 2, respectively, and let \( r_i = r_{1i} + r_{2i} \) be the number of subjects at risk in the combined sample. Analogously, let \( d_{1i}, d_{2i}, \) and \( d_i = d_{1i} + d_{2i} \) be the number of events at time \( t_i \).

Then, if \( H_0 : h_1(t) = h_2(t) \) is true, \( d_{1i} \) has a hypergeometric distribution with parameters \( (r_i, d_i, r_{1i}) \).

<table>
<thead>
<tr>
<th>Event</th>
<th>No event</th>
<th>At risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1</td>
<td>( d_{1i} )</td>
<td>( r_{1i} - d_{1i} )</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>( d_{2i} )</td>
<td>( r_{2i} - d_{2i} )</td>
</tr>
<tr>
<td>Merged</td>
<td>( d_i )</td>
<td>( r_i - d_i )</td>
</tr>
</tbody>
</table>

Since \( d_{1i} \sim \mathcal{HG}(r_i, d_i, r_{1i}) \), the expectation and variance of \( d_{1i} \) are

\[
\begin{align*}
\mathbb{E}d_{1i} &= r_{1i} \frac{d_i}{r_i}, \\
\text{Var}(d_{1i}) &= r_{1i} \frac{1}{r_i} \left( 1 - \frac{r_{1i}}{r_i} \right) \left( \frac{r_i - d_i}{r_i - 1} \right) d_i.
\end{align*}
\]

Note that in the terminology of the Kaplan–Meier estimator, the number of subjects with no event at time \( t_i \) is equal to \( r_i - d_i = r_{i+1} + \ell_i \), where \( \ell_i \) is the number of subjects censored in the time interval \((t_{i-1}, t_i)\) and \( r_{i+1} \) is the number of subjects at risk at the beginning of the subsequent interval \((t_i, t_{i+1})\).

The test statistic for testing \( H_0 : h_1(t) = h_2(t) \) against the two-sided alternative \( H_1 : h_1(t) \neq h_2(t) \) is

\[
\chi^2 = \frac{\left( \sum_{i=1}^{D} (d_{1i} - \mathbb{E}(d_{1i})) \right)^2}{\sum_{i=1}^{D} \text{Var}(d_{1i})}, \quad (16.4)
\]

which has a \( \chi^2 \)-distribution with 1 degree of freedom. The continuity correction 0.5 can be added to the numerator of the \( \chi^2 \)-statistic as \( \left( | \sum_{i=1}^{D} (d_{1i} - \mathbb{E}(d_{1i})) | - 0.5 \right)^2 \) when the sample size is small. If the statistic is calculated as \( \text{chi2} \), then its large values are critical and the \( p \)-value of the test is equal to \( 1 - \text{chi2cdf(chi2,1)} \).

If the alternative is one-sided, \( H_1 : h_1(t) < h_2(t) \) or \( H_1 : h_1(t) > h_2(t) \), then the preferable statistic is
\[ Z = \frac{\sum_{i=1}^{D} (d_{1i} - \mathbb{E}(d_{1i}))}{\sqrt{\sum_{i=1}^{D} \text{Var}(d_{1i})}} \]

and the \( p \)-values are \( \text{normcdf}(Z) \) and \( 1 - \text{normcdf}(Z) \), respectively. A more general statistic is of the form

\[ Z = \frac{\sum_{i=1}^{D} W(t_i) (d_{1i} - \mathbb{E}(d_{1i}))}{\sqrt{\sum_{i=1}^{D} W^2(t_i) \text{Var}(d_{1i})}}, \]

where \( W(t_i) = 1 \) (as above), \( W(t_i) = r_i \) (Gehan’s statistic), and \( W(t_i) = \sqrt{r_i} \) (Tarone–Ware statistic).

**Example 16.11. Mantel’s Logrank Step-by-Step.** To illustrate the logrank test, we consider a simple example. Consider two trials A and B with outcomes 4.5, 7, 7, 8.9+, 9.1, and 6.1+, 7, 9, 10+.

At combined event times \( t_i \) the tables need to be formed. Organize the data as follows:

<table>
<thead>
<tr>
<th>Event</th>
<th>No event</th>
<th>At risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( d_{1i} )</td>
<td>( r_{1i} - d_{1i} )</td>
</tr>
<tr>
<td>B</td>
<td>( d_{2i} )</td>
<td>( r_{2i} - d_{2i} )</td>
</tr>
<tr>
<td></td>
<td>( d_i )</td>
<td>( r_i - d_i )</td>
</tr>
</tbody>
</table>

In the merged sample, there are four times with events: 4.5, 7, 9, and 9.1, leading to four tables. At time 4.5, five in group A are at risk, one with the event. In group B, four are at risk, none with the event. For time 7, four are at risk and two with the event in A, and three are at risk and one with the event in B. Likewise for times 9 and 9.1. The four tables are

| A | 1 4 5 | A | 2 2 4 | A | 0 1 1 | A | 1 0 | 1 |
| B | 0 4 4 | B | 1 2 3 | B | 1 1 2 | B | 0 1 | 1 |
|   | 1 8 | 9 | 3 4 | 7 | 1 2 | 3 | 1 1 | 2 |

Since \( d_{11} \sim \mathcal{H}(9,1,5) \), the expectation and variance of \( d_{11} \) are respectively \( \mathbb{E}d_{11} = 5 \times \frac{1}{9} = \frac{5}{9} \), and \( \text{Var}(d_{11}) = \frac{5}{9} (1 - \frac{5}{9}) \left( \frac{9 - 1}{9 - 1} \right) \times 1 = \frac{20}{81} \). Thus, observed events for group A, their expectations, and variances are:

| \( d_{1i} \) | 1 2 0 1 |
| \( \mathbb{E}d_{1i} \) | 5/9 12/7 1/3 1/2 |
| \( \text{Var}(d_{1i}) \) | 20/81 24/49 2/9 1/4 |

Equation in (16.4) leads to \( \chi^2 = 0.6653 \). Under the null hypothesis this statistic has a \( \chi^2 \)-distribution with one degree of freedom, so the \( p \)-value is 0.4147 (\( 1 - \text{chi2cdf}(0.6653,1) \)).

Notice that \( \text{manteltwo.m} \) produces the same result.
Example 16.12. Histiocytic Lymphoma. The data (from McKelvey et al., 1976; Armitage and Berry, 1994) given below are survival times (in days) since entry to a trial by patients with diffuse histiocytic lymphoma. Two cohorts of patients are considered: (1) with stage III and (2) with stage IV of the disease. The observations with + are censored.

<table>
<thead>
<tr>
<th>Stage</th>
<th>6</th>
<th>19</th>
<th>32</th>
<th>42</th>
<th>43+</th>
<th>94</th>
<th>126+</th>
<th>169+</th>
<th>207</th>
<th>211+</th>
</tr>
</thead>
<tbody>
<tr>
<td>III</td>
<td>227+</td>
<td>253</td>
<td>255+</td>
<td>270+</td>
<td>310+</td>
<td>316+</td>
<td>335+</td>
<td>346+</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage</th>
<th>4</th>
<th>6</th>
<th>10</th>
<th>11</th>
<th>11</th>
<th>11</th>
<th>13</th>
<th>17</th>
<th>20</th>
<th>20</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>22</td>
<td>24</td>
<td>24</td>
<td>29</td>
<td>30</td>
<td>30</td>
<td>31</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>41+</td>
<td>43+</td>
<td>45</td>
<td>46</td>
<td>50</td>
<td>56</td>
<td>61+</td>
<td>61+</td>
<td>63</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>82</td>
<td>85</td>
<td>88</td>
<td>89</td>
<td>90</td>
<td>93</td>
<td>104</td>
<td>110</td>
<td>134</td>
<td>137</td>
<td>160+</td>
</tr>
<tr>
<td></td>
<td>169</td>
<td>171</td>
<td>173</td>
<td>175</td>
<td>184</td>
<td>201</td>
<td>222</td>
<td>235+</td>
<td>247+</td>
<td>260+</td>
<td>284+</td>
</tr>
<tr>
<td></td>
<td>290+</td>
<td>291+</td>
<td>302+</td>
<td>304+</td>
<td>341+</td>
<td>345+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using a logrank test, we will assess the equality of the two survival curves. The function `manteltwo.m` calculates $\chi^2$ statistic and the corresponding $p$-value:

```matlab
load 'YourDataPath\limphoma.mat'
tA = limphoma.survive(limphoma.stage == 1);
cA = limphoma.censored(limphoma.stage == 1);
tB = limphoma.survive(limphoma.stage == 2);
cB = limphoma.censored(limphoma.stage == 2);
[p ch2] = manteltwo(tA, cA, tB, cB)
% p = 0.0096
% ch2 = 6.7097
```

The hypothesis of equality of survival curves in this case is rejected, $p$-value is 0.0096. Kaplan–Meier estimators of the two survival functions are shown in Figure 16.4. MATLAB Central contains several functions conducting logrank test. Function `logrank.m` (Cardillo, 2008) is an example.

16.4 The Cox Proportional Hazards Model

We often need to take into account that survival is influenced by one or more covariates, which may be categorical (e.g., the kind of treatment a...
patient received) or continuous (e.g., the patient’s age, weight, or drug dosage). For simple situations involving a single factor with just two values (e.g., drug versus placebo), we discussed a method for comparing the survival curves for the two groups of subjects. For more complicated situations, a regression-type model that incorporates the effect of each predictor on the shape of the survival curve is needed.

Assume that the log hazard for subject $i$ can be modeled via a linear relationship:

$$\log h(t, x_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i},$$

where $x_i = x_{1,i}, \ldots, x_{p,i}$ is $p$-dimensional vector of covariates associated with subject $i$. The Cox model assumes that $\beta_0$ is a log baseline hazard, $\log h_0(t) = \beta_0$, namely the log hazard for a “person” for whom all covariates are 0 (Cox, 1972; Cox and Oakes, 1984). Alternatively, we can set the baseline hazard to correspond to a typical person for whom all covariates are averages of covariates from all subjects in the study. For the Cox model,

$$\log h(t, x_i) = \log h_0(t) + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}$$

or, equivalently,

$$h(t, x_i) = h_0(t) \times \exp\{\beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}\} = h_0(t) \times \exp\{x_i'\beta\}.$$

Inclusion of an intercept would lead to nonidentifiability because
\[ h_0(t) \times \exp \{ x_i' \beta \} = (h_0(t)e^{-\alpha}) \times \exp \{ \alpha + x_i' \beta \}. \]

This allows for some freedom in choosing the baseline hazard.

For two subjects, \( i \) and \( j \), the ratio of hazard functions

\[ h(t, x_i) / h(t, x_j) = \exp \{ (x_i' - x_j') \beta \} \]

is free of \( t \), motivating the name proportional. Also, for a subject \( i \),

\[ S(t, x_i) = (S_0(t))^{\exp \{ x_i' \beta \}}, \]

where \( S_0(t) \) is the survival function corresponding to the baseline hazard \( h_0(t) \). This follows directly from (16.1) and \( H(t, x_i) = H_0(t) \exp \{ x_i' \beta \} \).

In MATLAB, \texttt{coxphfit} fits the Cox proportional hazards regression model, which relates survival times to predictor variables. The following example uses \texttt{coxphfit} to fit Cox’s proportional hazards model.

\textbf{Example 16.13. Mayo Clinic Trial in PBC.} Primary biliary cirrhosis (PBC) is a rare but fatal chronic liver disease of unknown cause, with a prevalence of about 1/20,000. The primary pathologic event appears to be the destruction of interlobular bile ducts, which may be mediated by immunologic mechanisms.

The PBC data set available at StatLib is an excerpt from the Mayo Clinic trial in PBC of the liver conducted between 1974 and 1984. From a total of 424 patients that met eligibility criteria, 312 PBC patients participated in the double-blind, randomized, placebo-controlled trial of the drug D-penicillamine. Details of the trial can be found in Markus et al. (1989).

Survival statuses were recorded for as many patients as possible until July 1986. By that date, 125 of the 312 patients had died and 187 were censored.

The variables contained in the data set \texttt{pbc.xls|dat} are described in the following table:

- \texttt{casen = pbc(:,1);} % case number 1-312
- \texttt{lived = pbc(:,2);} % days lived (from registration to study date)
- \texttt{indicatord = pbc(:,3);} % 0 censored, 1 death
- \texttt{treatment = pbc(:,4);} % 1 - D-Penicillamine, 2 - Placebo
- \texttt{age = pbc(:,5);} % age in years
- \texttt{gender = pbc(:,6);} % 0 male, 1 female
- \texttt{ascites = pbc(:,7);} % 0 no, 1 yes
- \texttt{hepatomegaly = pbc(:,8);} % 0 no, 1 yes
- \texttt{spiders = pbc(:,9);} % 0 no, 1 yes
- \texttt{edema = pbc(:,10);} % 0 no, 0.5 yes/no therapy, 1 yes/therapy
- \texttt{bilirubin = pbc(:,11);} % bilirubin [mg/dl]
- \texttt{cholesterol = pbc(:,12);} % cholesterol [mg/dl]
- \texttt{albumin = pbc(:,13);} % albumin [gm/dl]
- \texttt{ucopper = pbc(:,14);} % urine copper [mg/day]
- \texttt{aphosp = pbc(:,15);} % alkaline phosphatase [U/liter]
To illustrate the CPH model, in this example we selected four predictors and formed a design matrix $X$ as

$$X = [\text{treatment} \ \text{age} \ \text{gender} \ \text{edema}];$$

The treatment has two values, 1 for treatment by D-penicillamine and 2 for placebo. The variable edema takes three values: 0 if no edema is present, 0.5 when edema is present but no diuretic therapy was given or edema resolved with diuretic therapy, and 1 if edema is present despite administration of diuretic therapy.

The variable lived is the lifetime observed or censored, a censoring vector is $1$-indicatord, and a baseline hazard is taken to be a hazard for which all covariates are set to 0.

$$[b, \log L, H, \text{stats}] = \text{coxphfit}(X, \text{lived}, \ldots \ 'censoring', 1\text{-indicatord}, \ 'baseline', 0);$$

The output $H$ is a two-column matrix as a discretized cumulative hazard estimate. The first column of $H$ contains values from the vector lived, while the second column contains the estimated baseline cumulative hazard evaluated at lived.

To illustrate the model, we selected two subjects from the study to find survival curves corresponding to their covariates. Subject #100 is a 51-year-old male with no edema who received placebo while subject #275 is a 38-year-old female with no edema who received D-penicillamine treatment.

First, we find cumulative hazards at the mean values of predictors, as well as for subjects #100 and #275, as

$$H(t, \mathbf{x}) = H_0(t) \times \exp\{\beta_1 x_1 + \cdots + \beta_4 x_4\},$$

$$H(t, x_i) = H_0(t) \times \exp\{\beta_1 x_{1,i} + \cdots + \beta_4 x_{4,i}\}, \quad i = 100, 275.$$

Here the estimators of coefficients $\beta_1, \ldots, \beta_4$ are

$$b'$$

| b' | 0.0831 | 0.0324 | -0.3940 | 2.2424 |
Note that the *treatment* coefficient \(0.0831 > 0\) indicates that, given all other covariates fixed, the placebo increases the risk over the treatment. Also note that *age* and *edema* statuses also increase the risk, while the risk for female subjects is smaller.

Next, from cumulative hazards we find survival functions

\[
\begin{align*}
S_{\text{mean}} &= \exp(-H_{\text{mean}}(:,2)); \\
S_{\text{subj}100} &= \exp(-H_{\text{subj}100}(:,2)); \\
S_{\text{subj}275} &= \exp(-H_{\text{subj}275}(:,2));
\end{align*}
\]

The subsequent commands plot the survival curves for an “average” subject (blue), as well as for subjects #100 (black) and #275 (red); see Figure 16.5.

```matlab
stairs(H(:,1),Smean,'b-','linewidth',2)
hold on
stairs(H(:,1),Ssubj100,'k-')
stairs(H(:,1),Ssubj275,'r-')
xlabel('$t$ (days)','$Interpreter','LaTeX')
ylabel('$\hat{S}(t)$','$Interpreter','LaTeX')
legend('average subject','subject #100', 'subject #275', 3)
axis tight
```

*Fig. 16.5* Cox model for survival curves for a “subject” with average covariates, subject #100 (51-year-old male on placebo, no edema), and subject #275 (38-year-old female on treatment, no edema).
16.5 Bayesian Approach

We will focus on parametric models in which the lifetime distributions are specified up to unknown parameters. The unknown parameters will be assigned prior distributions and the inference will proceed in a Bayesian fashion. Nonparametric Bayesian modeling of survival data is possible; however, the methodology is advanced and beyond the scope of this text. For a comprehensive coverage, see Ibrahim et al. (2001).

Let survival time $T$ have distribution $f(t|\theta)$, where $\theta$ is unknown parameter. For $t_1, \ldots, t_k$ observed and $t_{k+1}, \ldots, t_n$ censored times, the likelihood is

$$L(\theta|t_1, \ldots, t_n) = \prod_{i=1}^{k} f(t_i|\theta) \times \prod_{i=k+1}^{n} S(t_i|\theta).$$

If the prior on $\theta$ is $\pi(\theta)$, then the posterior is

$$\pi(\theta|t_1, \ldots, t_n) \propto L(\theta|t_1, \ldots, t_n) \times \pi(\theta).$$

The Bayesian estimator of hazard is

$$\hat{h}_B(t) = \int h(t|\theta)\pi(\theta|t_1, \ldots, t_n)d\theta$$

and the survival function is

$$\hat{S}_B(t) = \int S(t|\theta)\pi(\theta|t_1, \ldots, t_n)d\theta.$$

Example 16.14. Exponential Lifetimes with Gamma Prior. In Example 16.6 we found that for the exponential lifetime in the presence of censoring, the likelihood was $L(\lambda) = \lambda^k \exp\{-\lambda \sum t_i\}$, where $k$ is the number of uncensored data and $\sum_{i=1}^{n} t_i$ is the sum of all observed and censored times. The resulting MLE for $\lambda$ was

$$\hat{\lambda} = k / \sum_{i=1}^{n} t_i.$$

If a gamma $Ga(\alpha, \beta)$ prior on $\lambda$ is adopted, $\pi(\lambda) \propto \lambda^{a-1} \exp\{-\beta \lambda\}$, then the conjugacy of the exponential/gamma pair (page 341) leads to the posterior

$$\pi(\lambda|t_1, \ldots, t_n) \propto \lambda^{k+a-1} \exp\{-\beta + \sum_{i=1}^{n} t_i \lambda\},$$

from which the Bayes estimator of $\lambda$ is the posterior mean,
The posterior predictive distribution of future failure time $t_{n+1}$ is

$$f(t_{n+1}|t_1, \ldots, t_n) = \int_0^\infty \lambda e^{-\lambda t_{n+1}} \times \pi(\lambda|t_1, \ldots, t_n) d\lambda$$

$$= \frac{(k + \alpha)(\beta + \sum_{i=1}^n t_i)^{a+k}}{(\beta + \sum_{i=1}^n t_i + t_{n+1})^{a+k+1}}, \quad t_{n+1} > 0.$$ 

This distribution is known as an inverse beta distribution.

The Bayes estimator of hazard function coincides with the Bayes estimator of $\lambda$,

$$\hat{h}_B(t) = \frac{k + \alpha}{\beta + \sum_{i=1}^n t_i},$$

while the Bayes estimator of the survival function is

$$\hat{S}_B(t) = \left(1 + \frac{t}{\beta + \sum_{i=1}^n t_i}\right)^{-(k+\alpha)}.$$ 

The expression for $\hat{S}_B(t)$ can be derived from the moment-generating function of a gamma distribution.

When the posterior distribution is intractable, one can use WinBUGS.

### 16.6 Survival Analysis in WinBUGS

WinBUGS uses two arrays to define censored observations: observed (uncensored) times and censored times. For example, an input such as

```r
list(times = c(0.5, NA, 1, 2, 6, NA, NA),
t.censored = c(0, 0.9, 0, 0, 0, 9, 12))
```

corresponds to times \{0.5, 0.9+, 1, 2, 6, 9+, 12+\}.

In WinBUGS, direct time-to-event modeling is possible with exponential, Weibull, gamma, and log-normal densities.

There is a multiplier $I$ that is used to implement censoring. For example, if Weibull $dweib(r, mu)$ observations are on the input, the multiplier is an indicator that the observation exceeded time $t.censored[i]$:

```r
t[i] ~ dweib(r, mu) I(t.censored[i],)
```
16.6 Survival Analysis in WinBUGS

For uncensored data one sets \( t_{\text{censored}}[i] = 0 \). The above describes right-censoring. Left-censored observations are modeled using the multiplier \( I(t_{\text{censored}}[i]) \).

**Example 16.15. Bayesian Immunoperoxidase and BC.** In Example 16.7 MLEs and confidence intervals on \( \lambda_1 \) and \( \lambda_2 \) were found. In this example we find Bayes estimators and credible sets.

Before discussing the results, note that when observations are censored, their values are unknown parameters in the Bayesian model and predictions can be found. Because they are treated as parameters, all censored observations need to be initialized, and the initial values should exceed the censoring times. Here is the WinBUGS program that estimates \( \lambda_1 \) and \( \lambda_2 \).

```winbugs
model{
  for(i in 1:n1) {
    ImmPeroxNeg[i] ~ dexp(lam1) I(CensorIPN[i],)
  }
  for(i in 1:n2) {
    ImmPeroxPos[i] ~ dexp(lam2) I(CensorIPP[i],)
  }
  lam1 ~ dgamma(0.001, 0.001)
  lam2 ~ dgamma(0.001, 0.001)
}

DATA

list( n1 = 36, n2 = 9,
  ImmPeroxNeg=c(19, 25, 30, 34, 37, 46, 47, 51,
    56, 57, 61, 66, 67, 74, 78, 86,
    NA, NA, NA, NA, NA, NA, NA, NA,
    NA, NA, NA, NA, NA, NA, NA, NA, NA),
  CensorIPN = c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
    122, 123, 130, 130, 133, 134, 136, 141, 143, 148,
    151, 152, 153, 154, 156, 162, 164, 165, 182, 189),
  ImmPeroxPos = c(22, 23, 38, 42, 73, 77, 89, 115, NA),
  CensorIPP= c(0,0,0,0,0,0,0,0,0,0,0,144))

INITS

list(lam1=1, lam2 = 1,
  ImmPeroxNeg=c(NA, NA, NA, NA, NA, NA, NA, NA,
    NA, NA, NA, NA, NA, NA, NA, NA),
  ImmPeroxPos = c(NA, NA, NA, NA, NA, NA, NA, NA,
    NA, NA, NA, NA, NA, NA, NA, NA) )
```
The Bayes estimator of $\lambda_1$ is $\hat{\lambda}_{1,.B} = 0.004211$, and the 95% credible set is $[0.002404, 0.006505]$. The Bayes estimator and interval are close to their classical counterparts (Exercise 16.2).

Example 16.16. Smoking Cessation Experiment. The data set for this example comes from a clinical trial discussed in Banerjee and Carlin (2004). A number of smokers entered into a smoking cessation study, and 263 of them quit. These 263 quitters were monitored and checked to see if and when they relapsed. RelapseT is the time to relapse; it is either observed or censored, with the censoring indicator contained in the vector censored.time. The independent covariates are Age (age of the individual), AgeStart (age when he/she started smoking), SexF (Female=1, Male=0), SIUC (whether the individual received an intervention or not), and F10Cigs (the average number of cigarettes smoked per day).

A logistic distribution is constrained to a nonnegative domain to model RelapseT. The parameters of dlogis(mu,tau) are the mean mu, which depends on the linear combination of covariates, and tau, which is a rate parameter. The standard deviation is $\pi/(\sqrt{3}\tau) \approx 1.8138/\tau$.

```r
model {
  for (i in 1:N)
  {
    RelapseT[i] ~ dlogis(mu[i],tau) I(censored.time[i],)
  }
  for( j in 1:6){
    beta[j] ~ dnorm(0, 0.01)
  }
  tau ~ dgamma(1,0.01)
  meanT <- mean(mu[])
  sigma <- 1.8138/tau #1.8138 ~ pi/sqrt(3)
}
```

Evaluate Survival Curve for a Subject with covariates:
Ag <- 50; AgSt <- 18; SexF <- 0; S <- 1; Cigs <- 20;
for(i in 1:100) {
  time[i] <- i/10
  Surv[i] <- 1/(1 + exp(tau*(time[i] - fmu)))
}

# Data and Inits omitted (see Smoking.odc)

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>Merror</th>
<th>val2.5pc</th>
<th>median</th>
<th>val97.5pc</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta[1]</td>
<td>2.686</td>
<td>2.289</td>
<td>0.1254</td>
<td>-2.026</td>
<td>2.701</td>
<td>7.284</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>beta[2]</td>
<td>0.07817</td>
<td>0.03844</td>
<td>0.002078</td>
<td>0.009544</td>
<td>0.07815</td>
<td>0.1509</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>beta[3]</td>
<td>-0.07785</td>
<td>0.07222</td>
<td>0.003745</td>
<td>-0.2181</td>
<td>-0.07965</td>
<td>0.06802</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>beta[4]</td>
<td>-0.9555</td>
<td>0.5382</td>
<td>0.09515</td>
<td>-2.02</td>
<td>-0.9494</td>
<td>0.08981</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>beta[5]</td>
<td>1.666</td>
<td>0.5993</td>
<td>0.01482</td>
<td>0.4886</td>
<td>1.672</td>
<td>2.817</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>beta[6]</td>
<td>-0.02347</td>
<td>0.02459</td>
<td>0.00072</td>
<td>0.009544</td>
<td>0.07815</td>
<td>0.1509</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>meanT</td>
<td>5.314</td>
<td>0.384</td>
<td>0.00604</td>
<td>4.712</td>
<td>5.292</td>
<td>6.034</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>sigma</td>
<td>3.445</td>
<td>0.345</td>
<td>0.005367</td>
<td>2.841</td>
<td>3.421</td>
<td>4.187</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>tau</td>
<td>0.5317</td>
<td>0.05256</td>
<td>0.00072</td>
<td>0.009544</td>
<td>0.07815</td>
<td>0.1509</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>Surv[1]</td>
<td>0.9628</td>
<td>0.01129</td>
<td>0.009544</td>
<td>0.9336</td>
<td>0.9642</td>
<td>0.9808</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>Surv[2]</td>
<td>0.9609</td>
<td>0.01173</td>
<td>0.009544</td>
<td>0.9336</td>
<td>0.9623</td>
<td>0.9797</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>Surv[99]</td>
<td>0.1402</td>
<td>0.05249</td>
<td>0.00342</td>
<td>0.06123</td>
<td>0.1334</td>
<td>0.2645</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>Surv[100]</td>
<td>0.1342</td>
<td>0.05118</td>
<td>0.003311</td>
<td>0.05765</td>
<td>0.1274</td>
<td>0.2557</td>
<td>1001</td>
<td>100000</td>
</tr>
</tbody>
</table>

The **Surv** values (ordinate, mean, standard deviation, median, and quantiles) are exported to MATLAB as data file `smokingoutbugs.mat`. The file `smokingbugs.m` reads in the data and plots the posterior estimator of the survival curve (Figure 16.6). Note that the survival curve is \( S(t|\mu, \tau) = 1/(1 + \exp\{\tau(t - \mu)\}) \). The posterior distribution of \( S(t|\mu, \tau) \) is understood as a distribution of a function of \( \mu \) and \( \tau \) for \( t \) fixed.

![Fig. 16.6 Bayesian estimator of survival curve. The green bands are the 0.025 and 0.975 percentiles of the posterior distribution for \( S(t) \), while the blue errorbars have a size of posterior standard deviations.](image-url)
Example 16.17. **Duration of Remissions in Acute Leukemia.** A data set analyzed by Freireich et al. (1963), and subsequently by many authors, comes from a trial of 42 leukemia patients (under age of 20) treated in 11 US hospitals.

The effect of 6-mercaptopurine (6-MP) therapy on the duration of remissions induced by adrenal corticosteroids has been studied as a model for testing new agents. Some patients were treated with 6-MP and the rest were controls. The trial was designed as matched pairs. The matching was done with respect to remission status (partial = 1, complete = 2). Randomization to 6-MP or control arms was done within a pair. Patients were followed until leukemia relapsed or until the end of the study. Overall survival was not significantly different for the two treatment programs since patients maintained on placebo were treated with 6-MP when relapse occurred.

<table>
<thead>
<tr>
<th>Status</th>
<th>Contr 6-MP</th>
<th>Status</th>
<th>Contr 6-MP</th>
<th>Status</th>
<th>Contr 6-MP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>22</td>
<td>7</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>32+</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>12</td>
<td>23</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>8</td>
<td>22</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>17</td>
<td>6</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>16</td>
<td>14</td>
<td>2</td>
</tr>
</tbody>
</table>

**# Duration of Steroid-induced Remissions in Acute Leukemia**

```r
model {
  for (i in 1:n) {
    log(mu[i]) <- b0 + b1[treat[i]] + b2[status[i]]
    t[i] ~ dweib(r, mu[i]) I(t.cen[i],)
    S[i] <- exp(-mu[i] * pow(t[i], r));
    f[i] <- mu[i] * r * pow(t[i], r-1) * S[i]
    Lik[i] <- pow(f[i], 1 - delta[i]) * pow(S[i], delta[i]);
    logLik[i] <- log(Lik[i])
  }

  b0 ~ dnorm(0, 0.00001)
  b1[1] <- 0
e1[2] ~ dnorm(0, 0.001)
  b1[1] <- 0
  b1[2] ~ dnorm(0, 0.001)
  r ~ dgamma(0.01, 0.01)

  Dev <- -2 * sum(logLik[]) # deviance
}
```

**DATA**

```r
list(n = 42,
t = c(1, 22, 3, 12, 8, 17, 2, 11, 8, 12,
    2, 5, 4, 15, 8, 23, 5, 11, 4, 1, 8,}
```
10, 7, NA, 23, 22, 6, 16, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
NA, NA, NA, 6, NA, NA, 6, 13, NA, NA, NA,
\[
\begin{align*}
\beta[i] & \sim \text{dnorm}(0.0, 0.001) \\
\text{median}[i] & \leftarrow \text{pow}(\log(2) \times \exp(-\beta[i]), 1/r)
\end{align*}
\]

\[
\begin{align*}
r & \sim \text{dexp}(0.001)
\end{align*}
\]

DATA

```r
list( t = structure(.Data = c(12, 1, 21, 25, 11, 26, 27, 30, 13,
12, 21, 20, 23, 25, 23, 29, 35, NA, 31, 36, 32, 27, 23, 12,
18, NA, NA, 38, 29, 30, NA, 32, NA, NA, NA, NA, 25, 30,
37, 27, 22, 26, NA, 28, 19, 15, 12, 35, 35, 10, 22, 18, NA,
12, NA, NA, 31, 24, 37, 29, 27, 18, 22, 13, 18, 29, 28, NA,
16, 22, 26, 19, NA, NA, 17, 28, 26, 12, 17, 26), .Dim = c(4, 20)),
t.cen = structure(.Data = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 40, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
40, 40, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
40, 40, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
24, 0, 40, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
40, 40, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
29, 20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
10, 0, 0, 0, 0, 0)), .Dim = c(4, 20)), M = 4, N = 20)
```

INITS

```r
list( r=1, beta=c(0,0,0,0) ) #generate the rest
```

<table>
<thead>
<tr>
<th>mean</th>
<th>sd</th>
<th>MCerror</th>
<th>val2.5pc</th>
<th>median</th>
<th>val97.5pc</th>
<th>start sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>median[1]</td>
<td>23.82</td>
<td>1.977</td>
<td>0.02301</td>
<td>20.20</td>
<td>23.74</td>
<td>1001 1000000</td>
</tr>
<tr>
<td>median[2]</td>
<td>35.11</td>
<td>3.459</td>
<td>0.01650</td>
<td>29.23</td>
<td>34.79</td>
<td>1001 1000000</td>
</tr>
<tr>
<td>median[3]</td>
<td>26.80</td>
<td>2.401</td>
<td>0.02169</td>
<td>22.46</td>
<td>26.66</td>
<td>1001 1000000</td>
</tr>
</tbody>
</table>

Fig. 16.7 Posterior densities of median survival times median[1]–median[4].
16.7 Exercises


```matlab
y = exprnd(10, 50, 1); % Random failure times exponential(10)
d = exprnd(20, 50, 1); % Drop-out times exponential(20)
t = min(y, d); % Observe the minimum of these times
censored = (y > d); % Observe whether the subject failed
```

Using MATLAB’s `ecdf` calculate and plot empirical CDF and confidence bounds for arguments `t` and `censored`.

16.2. Immunoperoxidase. In the context of Example 16.7 find a confidence interval for \( \lambda_1 \), the rate parameter for immunoperoxidase-negative patients. Use the fact that for MLE \( \hat{\lambda}_1 \),

\[
\frac{\hat{\lambda}_1 - \lambda_1}{\hat{\lambda}/\sqrt{k_1}} \quad \text{or} \quad \sqrt{k_1} (\log \hat{\lambda}_1 - \log \lambda_1),
\]

both have an approximately standard normal distribution. Here, \( k_1 \) is the number of non-censored life-times.

16.3. Expected Lifetime. Let \( T \) be a lifetime with survival function \( S(t) \). Using integration by parts in the definition of \( \mathbb{E}T \), show

\[
\mathbb{E}T = \int_0^\infty S(t)dt.
\]

16.4. Rayleigh Lifetimes. A lifetime \( T \) would have Rayleigh distribution with scale parameter \( \theta \) if its CDF is given by

\[
F(t) = 1 - \exp \left\{ -\frac{t^2}{2\theta} \right\}, \ t \geq 0.
\]

(a) Find the PDF, survival function \( S(t) \), hazard \( h(t) \), and cumulative hazard \( H(t) \).
(b) Assume that lifetimes \( T_1 = 10, T_2 = 8, T_3 = 6 \) and \( T_4 = 10 \) have been observed. Find the MLE of \( \theta \). If \( ET^2 = 2\theta \), show that the corresponding moment matching estimator coincides with MLE. Evaluate the MLE for the four observed values.
(c) Find the theoretical median life \( t_{med} \). Is it close to observed median life \( T_{med} = (8 + 10)/2 = 9 \)?

16.5. Log-logistic Lifetimes. A lifetime \( T \) would have log-logistic \( \mathcal{LL}(\alpha, \beta) \) distribution if its PDF and CDF are given by
\begin{align*}
  f(t|\alpha, \beta) &= \frac{\alpha}{\beta} \left( \frac{t}{\beta} \right)^{\alpha-1} \left( 1 + \left( \frac{t}{\beta} \right)^{\alpha} \right)^{-\alpha}, \quad t \geq 0, \\
  F(t|\alpha, \beta) &= 1 - \frac{1}{1 + \left( \frac{t}{\beta} \right)^{\alpha}} = \frac{t^\alpha}{\beta^\alpha + t^\alpha}, \quad t \geq 0.
\end{align*}

Here the parameters \( \alpha \) (shape) and \( \beta \) (scale) are both positive.

If \( T \) has a log-logistic distribution with shape parameter \( \alpha \) and scale parameter \( \beta \), then \( Y = \log(T) \) has a logistic distribution with location parameter \( \log(\beta) \) and scale parameter \( 1/\alpha \).

The \( k \)th raw moment of \( T \) is given by

\[ E T^k = \beta^k \frac{k\pi/\alpha}{\sin(k\pi/\alpha)}, \quad k < \alpha. \]

(a) Find the survival function \( S(t) \), hazard \( h(t) \), and cumulative hazard \( H(t) \).

(b) Assume that \( \alpha = 2 \) and that lifetimes \( T_1 = 4, T_2 = 2, T_3 = 2, \) and \( T_4 = 7 \) from \( \mathcal{LL}(2, \beta) \) have been observed. Find the moment matching estimator of \( \beta \). Evaluate the estimator for the four observed values.

(c) Find \( t_p \), the \( p \)-quantile for \( T \). Estimate the median life \( t_{1/2} \) and quartiles \( t_{1/4} \) and \( t_{3/4} \) for the given data.

_HINT._ For cumulative hazard no need to integrate. Since in (b) \( \alpha = 2 \) only the first moments can be matched.


Using one of MATLAB’s built-in functions, _mle_ or _evfit_, fit the EV distribution to the data in Example 16.9. The PDF and CDF of EV distribution (Gumbel type I distribution for the minimum) are

\begin{align*}
  f(t) &= \frac{1}{b} \exp\left( \frac{t-a}{b} \right) \exp\left( -\exp\left( \frac{t-a}{b} \right) \right), \\
  F(t) &= 1 - \exp\left( -\exp\left( \frac{t-a}{b} \right) \right), \quad t \in \mathbb{R}.
\end{align*}

Plot the fitted survival function \( S(t|\hat{a}, \hat{b}) \) over the Kaplan–Meier estimator from Example 16.9. Comment on the agreement of the parametric and nonparametric estimators of the survival function.

_Hint:_ Estimators of \( a \) and \( b \) for EV distribution are _theta(1)_ and _theta(2)_ in
% Strength of Weathered Cord Data
T=[36.3, 41.7, ..., 41.9, 42.5];
censor=zeros(1,41),ones(1,7));
[theta, thetaCI] = mle(T, 'distribution', 'ev', 'censoring', censor)
% or: [theta thetaCI] = evfit(T,0.05,censor)

16.7. **Cumulative Hazards and Maxwell.** The components \( E_1, E_2, \) and \( E_3 \) are connected as in the graph below.

\[
\begin{array}{c}
E_1 \\
E_2 \\
E_3
\end{array}
\]

The components \( E_1 \) and \( E_2 \) have cumulative hazards

\[
H(t) = \begin{cases} 
\lambda t, & 0 \leq t < 1 \\
\lambda t^2, & t \geq 1
\end{cases}
\]

with \( \lambda_1 = 2 \) and \( \lambda_2 = 0.5 \), respectively. The component \( E_3 \) has a lifetime \( T_3 \) with the Maxwell distribution for which the PDF and CDF are

\[
f(t) = \sqrt{\frac{2}{\pi}} \frac{t^2}{\sigma^2} \exp \left\{ -\frac{t^2}{2\sigma^2} \right\}, \quad \text{and}
\]

\[
F(x) = 2\Phi \left( \frac{t}{\sigma} \right) - 1 - \sqrt{\frac{2}{\pi}} \frac{t}{\sigma} \exp \left\{ -\frac{t^2}{2\sigma^2} \right\}.
\]

The expectation of \( T_3 \) is \( \mathbb{E}T_3 = 2\sigma \sqrt{2/\pi} \). Unlike the \( \lambda \)'s, the parameter \( \sigma \) is not known in advance. From the previous experiments with components identical to \( E_3 \), the following survival times are available: 1, 3, 4, 1, and 1.

Using the above information, find the probability of the system being operational up to time (a) \( t = 0.5 \), (b) \( t = 2 \).

16.8. **A System with Log-Burr Lifetimes.** In the network from Exercise 16.7 all three components are identical and have a lifetime \( T \) given by the log-Burr CDF:

\[
F(t) = \frac{(1 - e^{\lambda t})^2}{1 + (1 - e^{\lambda t})^2} = 1 - \frac{1}{1 + (1 - e^{\lambda t})^2}, \quad t \geq 0, \lambda > 0.
\]
The parameter $\lambda$ is not known; however, it could be estimated. In previous experiments involving $n = 6$ of such components the following lifetimes were obtained: $T_1 = 3, T_2 = 5, T_3 = 3, T_4 = 6, T_5 = 2,$ and $T_6 = 5.$

(a) If it is known that for this distribution

$$E T = \frac{\pi}{4\lambda},$$

estimate $\lambda$ by the moment-matching approach.

(b) What is the probability that the network is operational at time $t = 4$?

16.9. **Censored Rayleigh.** The lifetime (in hours) of a certain sensor has Rayleigh distribution, with survival function

$$S(t) = \exp\left\{ -\frac{1}{2} \lambda t^2 \right\}, \quad \lambda > 0. $$

Twelve sensors are placed under test for 100 hours, and the following failure times are recorded 23, 40, 41, 67, 69, 72, 84, 84, 88, 100+, 100+. Here + denotes a censored time.

(a) If failure times $t_1, \ldots, t_r$ are observed, and $t_{r+1}^{+}, \ldots, t_n^{+}$ are censored, show that the MLE of $\lambda$ is

$$\hat{\lambda} = \frac{2r}{\sum_{i=1}^{r} t_i + \sum_{i=r+1}^{n} t_i^+}. $$

Evaluate the MLE for the given data. Consult Example 16.6.

(b) Calculate and plot the Kaplan–Meier estimator and superimpose $S(t)$ evaluated at $\hat{\lambda}$.

16.10. **MLE for Equally Censored Data.** A cohort of $n$ subjects is monitored in the time interval $[0, T]$, where $T$ is fixed in advance. Suppose that $r$ failures are observed ($r$ can be any number from 0 to $n$) at times $t_1, t_2, \ldots, t_r \leq T$. There are $(n - r)$ subjects that survived the entire period $[0, T]$, and their failure times are not observed.

Suppose that $f(t)$ is the density of a lifetime. The likelihood is

$$L = C \prod_{i=1}^{r} f(t_i) (1 - F(T))^{n-r}$$

for some normalizing constant $C$.

(a) Express the likelihood $L$ for the exponential lifetime distribution, that is, $f(t) = \lambda e^{-\lambda t}$, $t \geq 0$, and $F(t) = 1 - e^{-\lambda t}$, $t \geq 0$.

(b) Take the log of the likelihood, $\ell = \log(L)$.

(c) Find the derivative of $\ell$ with respect to $\lambda$. Set the derivative to 0 and solve for $\lambda$. If the solution $\hat{\lambda} \maximizes \ell$, $(\ell''(\hat{\lambda}) < 0)$, then $\hat{\lambda}$ is the MLE of $\lambda$. 

16.7 Exercises

(d) Show that the MLE is
\[ \hat{\lambda}_{mle} = \frac{r}{\sum_{i=1}^{r} t_i + (n - r)T}. \]

(e) If in the interval [0, 8] four subjects failed at times \( t_1 = 2, t_2 = 5/2, \) \( t_3 = 4, \) and \( t_4 = 5, \) and two subjects survived without failure, find the estimator \( \lambda, \) assuming that the lifetime has an exponential distribution. Check your calculations with

```
data = [2 2.5 4 5 8 8];
cens = [0 0 0 0 1 1];
lamrec = mle(data, 'distribution','exponential','censoring',cens)
lammle = 1/lamrec %MATLAB uses reciprocal parametrization
```

(f) What is the MLE of \( \lambda \) if the unobserved failure times are ignored, that is, only four observed failure times, \( t_1, t_2, t_3, \) and \( t_4, \) are used?

16.11. **Cancer of Tongue.** Sickel-Santanello et al. (1988) analyze data on 80 males diagnosed with cancer of the tongue. Data are provided in the file `tongue.dat`. The columns in the dataset are as follows:
(i) Tumor DNA profile (1 = aneuploid tumor, 2 = diploid tumor);
(ii) Time to death or on-study time (in weeks); and
(iii) Censoring indicator (1 = censored, 0 = observed)

(a) Calculate and plot Kaplan–Meier estimators for the two types of tumor.

(b) Using Mantel’s logrank procedure `manteltwo.m`, test the hypothesis that the two survival functions do not differ significantly.

(c) Fit the Cox proportional hazard regression with tumor profile as the covariate. What is the 95% CI for the slope \( \beta \)? Compare the results with the conclusions from (b).

16.12. **Rayleigh and Bayes.** It was observed that in clinical studies dealing with cancer survival times follow Rayleigh distribution with pdf
\[ f(x) = 2\lambda te^{-\lambda t^2}, \ t \geq 0, \ \lambda > 0. \]

(a) Show that the hazard function is linear.

(b) Find the mean survival time as a function of \( \lambda \).

(c) For \( t_1, \ldots, t_k \) observed and \( t_{k+1}, \ldots, t_n \) censored times, show that the likelihood is proportional to
\[ L \propto \lambda^k \exp\{-\lambda \sum_{i=1}^{n} t_i^2\}. \]

If the prior on \( \lambda \) is gamma \( \mathcal{G}(\alpha, \beta) \), show that the posterior is gamma \( \mathcal{G}(\alpha + k, \beta + \sum_{i=1}^{n} t_i^2) \).
(d) Show that the Bayes estimators of the hazard and survival functions are

\[ \hat{h}_B(t) = \frac{2(k + \alpha)t}{\beta + \sum_{i=1}^{n} t_i^2} \quad \text{and} \quad \hat{S}_B(t) = \left(1 + \frac{t^2}{\beta + \sum_{i=1}^{n} t_i^2}\right)^{-(k+\alpha)}. \]

16.13. **Experiment at Strangeways Lab.** Pike (1966) provides data from an experiment conducted by Glücksmann and Cherry at Strangeways Laboratory, Cambridge, UK, in which two groups of female rats were exposed to carcinogen DMBA, and the number of days to death from vaginal cancer was recorded. The two groups were distinguished by pretreatment regime.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>143 164 188 188 190 192 206 209 213 216 216* 220 227 230 234 244* 246 265 304</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 2</td>
<td>142 156 163 198 204* 205 232 232 233 233 233 239 240 261 280 280 296 296 323 344*</td>
</tr>
</tbody>
</table>

The * on two entries in each group indicate that the observation was censored.

(a) Find and plot Kaplan–Meier estimators of survival functions for the two groups.

(b) Using the logrank procedure, test the hypothesis that the two survival functions are the same at significance level \( \alpha = 0.05 \).

16.14. **Exercise Stress Test.** Campbell and Swinscow (2009) describe an experiment in which 20 patients, 10 of normal weight and 10 severely overweight, underwent an exercise stress test. The patients had to lift a progressively increasing load for up to 12 minutes, but they were allowed to stop earlier if they could do no more. On two occasions the equipment failed before 12 minutes. The times (in minutes) achieved were:

<table>
<thead>
<tr>
<th>Normal weigh</th>
<th>2, 4, 6, 8, 8**, 9, 10, 12*, 12*, 12*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overweight</td>
<td>1, 3, 4, 5, 6, 7, 7**, 9, 11, 12*</td>
</tr>
</tbody>
</table>

Here * means that the end of test was reached, and ** stands for equipment failure.

Use the logrank test to compare these two groups. Report the \( p \)-value.

16.15. **Western White Clematis.** Muenchow (1986) tested whether male or female flowers (of *Clematis ligusticifolia*) were equally attractive to insects. The data in table represent waiting times (in minutes), which includes censored data (observations with +).
Compare survival functions for male and female flowers.

16.16. **Gastric Cancer Data.** Stablein et al. (1981) provide data on 90 patients affected by locally advanced, nonresectable gastric carcinoma. The patients are randomized to two treatments: chemotherapy alone (coded as 0) and chemotherapy plus radiation (coded as 1). Survival time is reported in days, with censoring indicator (1 = censored, 0 = death).

Data are provided in file: `gastric.cls|dat|xlsx` where the column 1 is the treatment code, column 2 is the survival time, and column 3 is censoring indicator:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Survival Time</th>
<th>Censoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1512</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1519</td>
<td>1</td>
</tr>
</tbody>
</table>

Is the treatment type significantly affecting the survival time? Use the log-rank test and Cox proportional hazards regression to answer this question.

16.17. **Duration of Remissions in Acute Leukemia.** In Example 16.17 times to remission in leukemia patients (controls and treated with 6-MP drug) were analyzed using a Bayesian model. For this data compare the two survival functions, for 6-MP treatment and control, using the log-rank test. Ignore matching by variable status. You can use the function `manteltwo.m`.

16.18. **Time to Second Birth.** The Medical Birth Registry of Norway was established in 1967 and contains information on all births in Norway since that time. The data set `secondbirth.dat|xlsx` distilled from the registry
provides the time between first and second births for a selection of 53,558 women.\footnote{The Medical Birth Registry of Norway is acknowledged for allowing the usage of the data and Dr. Stein Emil Vollset for providing the data.}

It is of interest to explore whether this time was possibly affected if the first child died within one year of its birth.

The data set contains the following variables (as columns)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Age of mother at first birth (in years)</td>
</tr>
<tr>
<td>sex</td>
<td>Sex of first child (1 = boy, 2 = girl)</td>
</tr>
<tr>
<td>death</td>
<td>First child died within one year (0 = no, 1 = yes)</td>
</tr>
<tr>
<td>time</td>
<td>Time from first birth to second birth or censoring (in days)</td>
</tr>
<tr>
<td>status</td>
<td>Censoring indicator (0 = censored, 1 = birth)</td>
</tr>
</tbody>
</table>

(a) Fit the Cox proportional hazard regression with variables age, sex, and death as covariates. What is the 95% CI for the parameter $\beta_3$ corresponding to variable death? Is the variable sex significant?

(b) Plot survival functions for death = 0 and death = 1.

(c) Using Mantel’s procedure \texttt{manteltwo.m}, test the hypothesis that the two survival functions are not significantly different.

16.19. \textbf{Rats on Three Diets.} The data is taken from the study by King et al. (1979). The researchers studied influence of dietary fat, food type and amount, and the dietary antioxidant butylated hydroxytoluene (BHT) on tumor induction and tumor growth by 7,12-dimethyl-benz[a]anthracene.

The study was to determine whether ingestion of polyunsaturated fat decreased or antagonized the inhibitory action of the antioxidant in comparison to diets that either contained equivalent amounts of a saturated fat or were very low in fat. The data provided in rat.csv consists of the tumor-free time (in days) in 90 rats on three different diets (column 1), censoring indicator (column 2), and a diet code (column 3). The three diets are coded according to the fat content: 1 stands for a low fat diet, 2 for a saturated fat diet, and 3 for an unsaturated fat diet.

(a) Fit the Weibull distribution for each of the three diets. Are the confidence intervals for the Weibull parameters overlapping?

(b) Find the Kaplan–Meier estimators for each of the three diets and superimpose the Weibull survival functions fitted in (a).

(c) Compare survival functions for the low fat and saturated fat diets. Use Mantel’s logrank test with $\alpha = 0.05$.

16.20. \textbf{Dukes‘ C Colorectal Cancer and Diet Treatment.} Colorectal cancer is a common cause of death. In the advanced stage of disease, when the disease is first diagnosed in many patients, surgery is the only treatment. Cytotoxic drugs, when given as an adjunct to surgery, do not prevent relapse and do not increase the survival in patients with advanced disease.
Interest has been shown, at least by patients, in a nutritional approach to treatment, where diet plays a critical role in the disease management program.

In a controlled clinical trial, McIllmurray and Turkie (1987) evaluated the diet treatment in patients with Dukes’ C colorectal cancer, because the residual tumour mass is small after operation, the relapse rate is high, and no other effective treatment is available. The diet treatment consisted of linolenic acid, an oil extract of the seed from the evening primrose plant *Onagraceae Oenothera biennis* and vitamin E.

The data for the treatment and control patients are given below:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Survival time (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linoleic</td>
<td>1+, 5+, 6, 6, 9+, 10, 10+, 12, 12, 12, 12+, 13+, 15+, (n₁ = 25)</td>
</tr>
<tr>
<td>Control</td>
<td>3+, 6, 6, 6, 8, 12, 12, 12+, 15+, 16+, 18+, 18+, 20, 22+, (n₂ = 24)</td>
</tr>
</tbody>
</table>

(a) Estimate the Weibull parameters for the two groups and superimpose the resulting \( \hat{S}(t) \) on the corresponding Kaplan–Meier estimators.

(b) For both treatment (linoleic acid) and control groups, find times at which 50% of the patients are surviving. Use \( \hat{S}(t) \) from (a).

(c) Using Mantel’s logrank procedure `manteltwo.m`, test the hypothesis that the two survival functions are the same.

A starter MATLAB file with data, `dukes.m`, is provided.

**MATLAB AND WINBUGS FILES AND DATA SETS USED IN THIS CHAPTER**

http://statbook.gatech.edu/Ch16.Survival/

- chada.m, cordband.m, coxreadmissions.m, ImmunoPerox.m, KMmuenchow.m, leukemia-direct.m, leukemia-remission.m, lifetables.m, limphomaLR.m, logrank.m, logrank1.m, MacDonaldCancer.m, matneltwo.m, Muenchow.m, PBC.m, ple.m, secondbrth.m, simulation1.m, simulation2.m, smokingbugs.m, tongue.m, weibullsim.m

- ibrahim1.odc, ibrahim2.odc, Immunoperoxidase.odc, Leukemia.odc, photocar.odc, Smoking.odc

- gehan.dat, KMmuenchow.txt, limphoma.mat, pbc.xls, pbcdatal.dat, prostatecan.dat, secondbirth.dat|xlsx, smokingoutbugs.mat, tongue.dat|xlsx
CHAPTER REFERENCES


